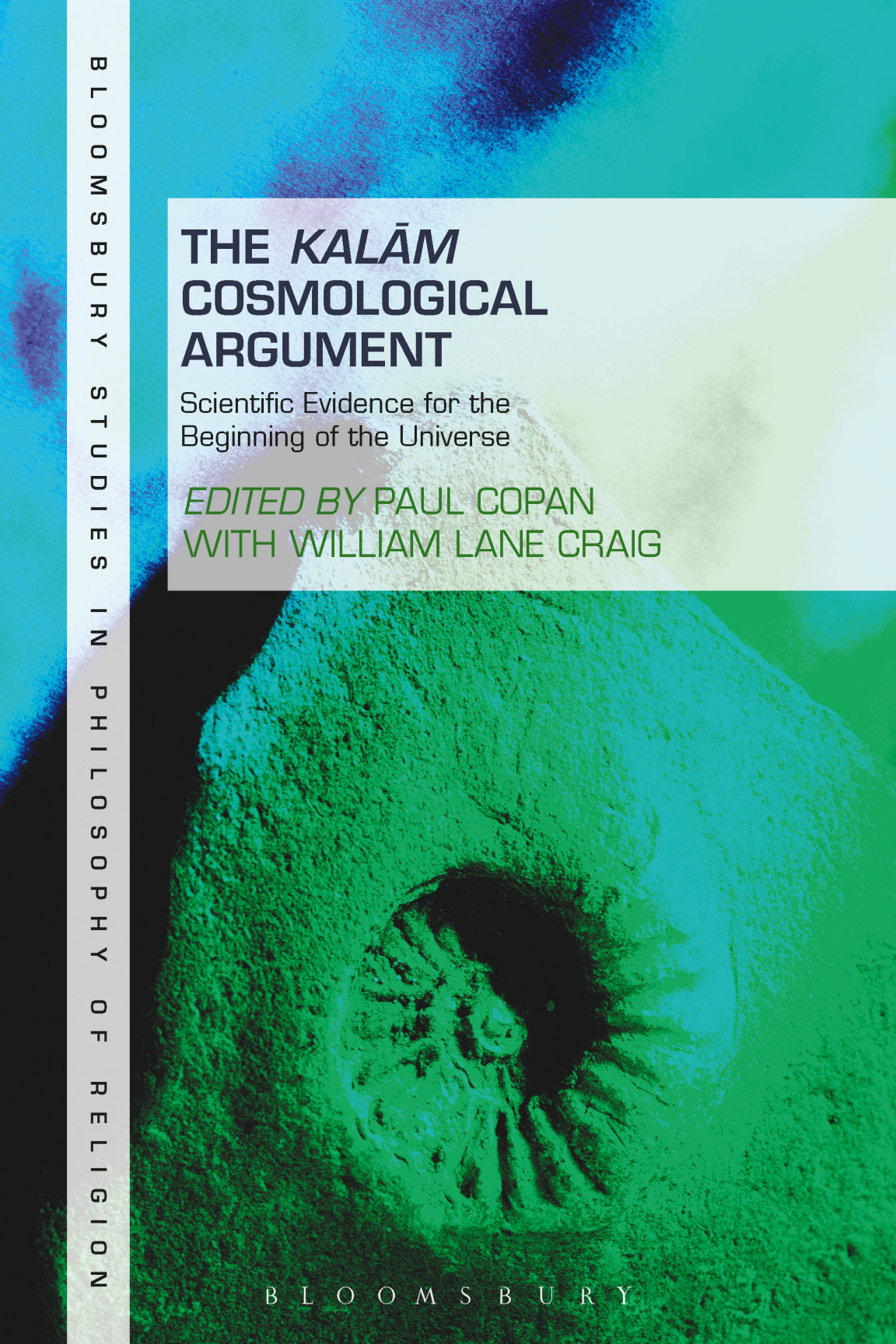
BLOOMSBURY STUDIES IN PHILOSOPHY OF RELIGION
THE KALĀM COSMOLOGICAL ARGUMENT
Scientific Evidence for the Beginning of the Universe
EDITED BY PAUL COPAN
WITH WILLIAM LANE CRAIG
BLOOMSBURY

The *Kalām* Cosmological Argument

The *Kalām* Cosmological Argument

Scientific Evidence for the Beginning of the Universe

Edited by Paul Copan
with William Lane Craig

Bloomsbury Academic
An imprint of Bloomsbury Publishing Inc

BLOOMSBURY
NEW YORK • LONDON • OXFORD • NEW DELHI • SYDNEY

Bloomsbury Academic
An imprint of Bloomsbury Publishing Inc

1385 Broadway
New York
NY 10018
USA

50 Bedford Square
London
WC1B 3DP
UK

www.bloomsbury.com

BLOOMSBURY and the Diana logo are trademarks of Bloomsbury Publishing Plc

First published 2018

© Paul Copan, William Lane Craig, and Contributors, 2018

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage or retrieval system, without prior permission in writing from the publishers.

No responsibility for loss caused to any individual or organization acting on or refraining from action as a result of the material in this publication can be accepted by Bloomsbury or the editors.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress.

ISBN: HB: 978-1-5013-3587-7
ePub: 978-1-5013-3588-4
ePDF: 978-1-5013-3589-1

Series: Bloomsbury Studies in Philosophy of Religion

Cover design: Louise Dugdale
Cover image © Getty Images/Corbis/VCG

Typeset by Deanta Global Publishing Services, Chennai, India

To find out more about our authors and books visit www.bloomsbury.com. Here you will find extracts, author interviews, details of forthcoming events, and the option to sign up for our newsletters.

Contents

Acknowledgments vi
Foreword vii

Introduction *Paul Copan* 1

Scientific Confirmation

2.3 The Inductive Argument Based on Expansion of the Universe

1 The *Kalām* Cosmological Argument: "Science" Excerpt *William Lane Craig and James D. Sinclair* 12
2 Why the Big Bang Singularity Does Not Help the *Kalām* Cosmological Argument for Theism *J. Brian Pitts* 80
3 On Non-Singular Spacetimes and the Beginning of the Universe *William Lane Craig and James D. Sinclair* 110
4 The Beginning of the Universe *Alexander Vilenkin* 150

2.4 The Inductive Argument Based on Thermodynamic Properties of the Universe

5 A Dying Universe: The Long-Term Fate and Evolution of Astrophysical Objects *Fred C. Adams and Gregory Laughlin* 159
6 Heat Death in Ancient and Modern Thermodynamics *Gábor Kutrovátz* 232
7 Entropy and Eschatology: A Comment on Kutrovátz's Paper "Heat Death in Ancient and Modern Thermodynamics" *Milan M. Ćirković* 243
8 The Generalized Second Law Implies a Quantum Singularity Theorem *Aron C. Wall* 252

Conclusion Therefore, the Universe Has a Cause

9 Libertarian Agency and the Craig/Grünbaum Debate about Theistic Explanation of the Initial Singularity *J. P. Moreland* 299
10 Causation and the Logical Impossibility of a Divine Cause *Quentin Smith* 312
11 Beyond the Big Bang *William Lane Craig* 333

Permissions 353
Index 356

Acknowledgments

I would like to express my gratitude to Haaris Naqvi at Bloomsbury for helping steer this volume through to publication. Thanks also to Stewart Goetz, a philosophy of religion editor for Bloomsbury, for recommending the proposal for publication. It has been a pleasure to work with Bloomsbury editor Katherine De Chant and Deanta Global's project manager Leeladevi Ulaganathan, both of whom have been gracious and resourceful. I very much appreciate Bloomsbury's production editor James Tupper for his behind-the-scenes labors. I am indebted to Mark Nowacki, Alexander Pruss, and James Sinclair, who offered helpful recommendations on articles to be included in this two-volume anthology. Special thanks to William Lane Craig, who offered wise counsel and useful suggestions during the anthology-production process. I am grateful to Aron Wall for working to get his revised article to me as well as Steven Wheeler, editor of *Inference*, for making special provisions on Alexander Vilenkin's essay. I appreciate the diligent work of David S. Oderberg, Andrew Loke, and Robert C. Koons, for making available original essays for the volumes of this anthology. Thanks to Tim Bayless and Lee Koz for their very practical assistance in the permissions process. I am grateful to my university—Palm Beach Atlantic University—for granting a sabbatical to devote time to writing and editing—and to Wycliffe Hall, Oxford University, for its invitation to be a visiting scholar there during the 2017 Hilary and Trinity terms. Finally, I also acknowledge indebtedness to my late mentor and friend, Stuart C. Hackett: *Requiescat in pace*.

Foreword

This is the second volume of an anthology on the medieval, but recently revived, *kalām* cosmological argument for God's existence. This argument contends that the series of past events is finite, that therefore the universe began to exist, and that the cause of the universe must be a personal one.

This argument has received fresh wind in its sails from two recent scientific discoveries that have yielded plausible prima facie physical evidence for the beginning of the universe. The expansion of the universe points to its beginning—to a Big Bang—as one retraces the universe's expansion in time. And the second law of thermodynamics, which implies that the universe's energy is progressively degrading, suggests that the universe began with an initial low entropy condition.

In contrast to the first volume of this anthology, this particular volume reviews and assesses the merits of the latest scientific evidences for the universe's beginning. The final portion of this book is a discussion of the *kalām* argument's conclusion—that the universe has a cause, indeed, a personal cause with properties of theological significance.

The Introduction to this volume not only presents an overview of the *kalām* cosmological argument, its background, and its recent resurgence (as does the Introduction in the companion volume); but it also provides chapter summaries of the essays on both the inductive (scientific) support for the second premise of the argument and the conclusion of the argument.

I trust that this work and its companion volume—*The Kalām Cosmological Argument: Philosophical Arguments for the Finitude of the Past*—will provide an engaging, fresh presentation of this enduring argument for God's existence from the vantage point of both its supporters and its opponents.

Introduction

Paul Copan

In 1979 William Lane Craig published *The Kalām Cosmological Argument*.[1] Like Hume's *Enquiry*, it fell stillborn from the presses, going immediately out of print. But over the decades since its first publication, it has steadily grown in influence until today; the *kalām* cosmological argument (KCA) has become one of the most widely debated arguments of natural theology. Quentin Smith, writing in *The Cambridge Companion to Atheism* (2006), observes,

> A count of the articles in the philosophy journals shows that more articles have been published about Craig's defense of the Kalam argument than have been published about any other philosopher's contemporary formulation of an argument for God's existence. Surprisingly, this even holds for Plantinga's ontological argument and Plantinga's argument that theism is a rationally acceptable basic belief. The fact that theists and atheists alike "cannot leave Craig's Kalam argument alone" suggests that it may be an argument of unusual philosophical interest or else has an attractive core of plausibility that keeps philosophers turning back to it and examining it again.[2]

Craig has continued to develop the argument of the original book (now back in print), revising, updating, and supplementing it and responding to published critiques.[3]

Craig's interest in the KCA (whose provenance will be explained momentarily) was sparked in 1971 by his reading of a little-known book by the Christian philosopher Stuart C. Hackett (1925–2012)[4] entitled *The Resurrection of Theism* (1957).[5] In his book Hackett called attention to this largely forgotten version of the cosmological argument.[6] Hackett argued that an infinite temporal regress of events—a series of past causes and effects—is impossible. Even though an infinite series may be conceivable in mathematics without contradiction, it involves "self-contradiction in the realm of being"—the realm of actual existents. He wrote:

> We are, in fact, familiar with the infinite series in mathematics: for example . . . -3, -2, -1, 0, 1, 2, 3 . . .; and we do not say that such a series involves self-contradiction. Why, then should we say that such a series does involve self-contradiction in the realm of being?
>
> I conceive that it will be a sufficient answer to this objection if I show that the hypothesis of an infinite series of actual existents involves self-contradiction.

> This is precisely the case, for either such a series consists of determinate parts, or it does not. If it does not, the series would have no applicability to the present universe, since, both spatially and temporally, the total manifest universe consists of determinate parts. But if the proposed series is thus constituted of such parts, self-contradiction is manifestly involved: for no series of determinate parts could add up to an actually infinite series. Yet such an actually infinite series must exist unless the series culminates in the absolutely self-conditioned being or God, which is contrary to the hypothesis that the series is actually infinite. Furthermore, suppose that there does exist such an infinite series of causes and effects. At any given point in time, the series would be infinite, yet greater than at any previous time: but this is self-contradictory since an actually infinite series does not admit of increase or decrease. Therefore, either the series is not infinite, or the progression of temporal events is an illusion; but then there is no way to account for the illusion. I therefore conclude that an infinite series of actual existents is inconceivable because self-contradictory and that therefore the objection is invalid. And if someone should deny that the law of contradiction applies here, I answer that this type of objection is self-annihilating.[7]

What Hackett called a "self-contradiction in the realm of being" would today be called a metaphysical impossibility. Craig was particularly intrigued by the notion that an infinite regress of past events, while not strictly logically impossible, might be metaphysically impossible.

As background for engagement with Hackett's argument, Craig undertook to study the history of the cosmological argument for the existence of God. As a result, he proposed a threefold taxonomy of cosmological arguments which has been widely adopted: the Thomistic cosmological argument, the Leibnizian cosmological argument, and the *kalām* cosmological argument.[8]

Versions of the cosmological argument

Unlike the monistic Advaita Vedanta school of Hinduism, which denies the existence of an external world (as *maya*, "illusion") and affirms the pure consciousness of Brahman as the sole reality, cosmological arguments take for granted the fact of the universe's existence. Furthermore, they aim to show there must exist—or likely exists—a First Cause or Sufficient Reason for the existence of the world or universe. The argument concludes that a Cause must ultimately be responsible for contingency, change, motion, causation within the universe or with reference to the existence of the universe (*kosmos*) itself. Besides Plato and Aristotle, other notables who have defended the cosmological argument include John Philoponus, ibn Sīna, al-Ghāzālī, Maimonides, Anselm, Aquinas, Bonaventure, Scotus, Descartes, Spinoza, Leibniz, and Locke.

The *Thomistic* cosmological argument is named for the medieval philosopher-theologian Thomas Aquinas. Yet it was Aristotle who supplied Thomas with the metaphysical structure that inspired his formulation of the "Five Ways," or, purported "proofs" for God's existence. The first three ways (respectively) argue for God from

motion, the causation of existence, and contingent and necessary entities. His is not an argument for a temporal beginning to the universe, which Aquinas thought could not be demonstrated or proven (Aquinas rejected mere probability arguments). In fact, the universe's beginning is like the doctrine of the Trinity—a matter specially revealed in Scripture and accepted by faith.

Rather, Aquinas's is a kind of metaphysical argument. God, who has no potentiality but is pure actuality, must exist in order to bring potentiality to actuality. Moreover, only a necessary, uncaused being, whose essence and existence are identical—that which is being (*esse*) itself—could be the ground for the existence of all contingent things that participate in or have being (*habens esse*).

Following Aristotle's potency-act distinction, Aquinas refers to *motion*, by which he means bringing about something that is, say, potentially hot to being actually hot. Ultimately, change must come from something unchangeable and without potentiality—namely, God. God is *ipsum esse subsistens*—subsistent being itself, a being whose essence just *is* existence. Thus, an infinite series of what he called essentially ordered causes is impossible (e.g., a hammer that is made to pound a nail, which is moved by an arm, which is moved by the choice of a personal agent, and so on). Or consider the movement of a train: the simultaneous movement of the cars cannot be explained by the movement of the preceding car, but by the engine at the front. Likewise, God, the ultimate Cause and Unmoved Mover, is responsible for the actualization of the potential: "this all men speak of as God" and "this we call God," Aquinas concluded in his *Summa Theologica* (I.2.3).[9]

Though superficially similar, the *Leibnizian* cosmological argument is crucially different from the Thomistic argument. Without grounding his argument in Aristotelian-Thomistic metaphysics, Leibniz started with the "first question which should rightly be asked"—namely, "Why is there something rather than nothing?"[10] To answer this question, he appealed to the Principle of Sufficient Reason: "No fact can be real or existent, no statement true, unless there be a sufficient reason why it is so and not otherwise,"[11]

This is in stark contrast to Bertrand Russell's assertion that the existence of the universe is a brute fact: "The universe is just there, and that's all."[12] Leibniz rejected such an idea—not just for the universe but also for God. Contingent things exist by virtue of something independent of them. For example, all of the constituents or states of the universe are not self-explanatory, since they are contingent. An infinite regress of contingent events still does not account for the *entire* series of events—or all of the contingent entities comprising the universe. Simply stringing together contingent things does not thereby render them necessary. This would be like check-kiting: a person with no money in the bank writes Check A to take care of his pressing debts and then Check B to cover Check A, and so on. By contrast, God, a necessary being, is self-sufficient and self-explanatory, needing nothing outside himself to account for his existence. Indeed, necessary existence is bound up within his nature: as the greatest conceivable being, God cannot not exist. So in addition to being the sufficient reason for all contingent entities, God, by virtue of his own nature, is the sufficient reason for his own existence.

Though initially skeptical of Leibniz's version of the argument, Craig eventually came to formulate and defend a Leibnizian argument along the following lines:

I. Anything that exists has an explanation of its existence, either in the necessity of its own nature or in an external cause.
II. If the universe has an explanation of its existence, that explanation is God.
III. The universe exists.
IV. Therefore, the explanation for the universe's existence is God.[13]

But that is a story for another day; our interest lies in the third type of cosmological argument identified by Craig.

The KCA is the version of the cosmological argument that Craig found to lie behind Hackett's argument. "*Kalām,*" Arabic for "word," came to denominate the philosophical theology pursued by medieval Islamic scholastics. Proponents of this version of the argument, going all the way back to the Aristotelian commentator John Philoponus in pre-Islamic Alexandria, argued against the possibility of an infinite temporal regress of events—they held, rather, the universe came into being a finite time ago. The argument aims to show that an infinite temporal regress of events is impossible and that therefore the universe began to exist without a material cause. In conjunction with the causal principle that whatever begins to exist has a cause, this fact implies that the universe has a transcendent (efficient) cause of its existence.

The KCA

The Muslim philosopher al-Ghāzālī presented the KCA in this simple form: "Every being which begins has a cause for its beginning; now the world is a being which begins; therefore, it possesses a cause for its beginning."[14] Ghāzālī's argument is disarmingly simple:

1. Whatever begins to exist has a cause of its beginning.
2. The universe began to exist.
3. Therefore, the universe has a cause of its beginning.

If all goes well, once one has arrived at a transcendent cause of the beginning of the universe, a conceptual analysis of what it is to be a cause of the universe's beginning enables one to recover a number of striking attributes of such a cause, including timelessness, immateriality, and, ultimately, personhood.

Craig initially adopted Ghāzālī's formulation but more recently has observed that the argument's success does not depend on a premise so sweeping as (1), rather the more modest

1'. If the universe began to exist, the universe has a cause of its beginning.

will suffice and is even more perspicuously true. Although Craig has offered various arguments in defense of the causal premise, his interest clearly lies in (2). He has offered both deductive and inductive arguments in support of the key claim that the universe began to exist.

Deductive arguments

2.1 The deductive argument based on the impossibility of an actual infinite:
  2.11 An actual infinite cannot exist.
  2.12 An infinite temporal regress of events is an actual infinite.
  2.13 Therefore, an infinite temporal regress cannot exist.

2.2 The deductive argument based on the impossibility of the formation of an actual infinite by successive addition:
  2.21 A collection formed by successive addition cannot be actually infinite.
  2.22 The temporal series of past events is a collection formed by successive addition.
  2.23 Therefore, the temporal series of past events cannot be actually infinite.

Inductive arguments

2.3 The inductive argument based on the expansion of the universe
2.4 The inductive argument based on the thermodynamic properties of the universe

The first deductive argument is an expression of finitism. On Craig's Aristotelian view, the only infinite which can actually exist is the potential infinite, which is merely a limit concept, not a number. There are no examples of actual infinites in the concrete world, and the notion of the potential infinite suffices for the pursuit of modern science. But what about the abstract mathematical realm described by infinite set theory? Here the proponents of the KCA have either made an in principle distinction between causally effete abstract objects and concrete objects, holding that the absurdities attending the existence of an actual infinite apply only to causally connected objects, or they, like Craig himself, adopt an anti-realist view of such abstract objects, denying them existence in any metaphysically heavy sense.[15]

The second deductive argument serves to bring to the fore the theory of time, which underlies the KCA, namely, a tensed, as opposed to tenseless, theory of time which affirms the objective reality of temporal becoming. Premise (2.22) presupposes that the series of temporal events does not exist tenselessly but comes into being *seriatim*. Therefore, the present event is reached by a genuine lapse of an actually infinite number of prior events one after another, if the universe is beginningless. This feature of an infinite past occasions manifold difficulties, famously exemplified in the thesis of Kant's First Antinomy concerning time. Lest one be tempted to avoid these difficulties by opting for a tenseless theory of time, Craig has argued at length in a pair of volumes for a tensed theory of time and against a tenseless theory.[16]

What is especially intriguing about the KCA's claim of the beginning of the universe is the scientific evidence to which Craig appeals from relatively recent discoveries pointing to the universe's temporal finitude. As indicated above, the two major inductive arguments supporting the second premise of the KCA are based on, first, the expansion of the universe and, second, the thermodynamic properties of the universe.

Consider, first, the evidence from the universe's expansion. In the 1920s, Alexander Friedmann and Georges LeMaître independently proposed an expanding universe and an edge to space-time. Their space-time model of a four-dimensional universe resembles a cone with a hard edge or boundary to the universe's beginning—and that of space, time, matter, and energy.[17] The standard Big Bang cosmological model describes the expansion of space itself—not simply matter within the universe. As we "reverse engineer" the universe and backtrack its trajectory toward zero, we arrive at a singularity—a boundary to space-time.

This beginning and expansion were confirmed by astronomer Edwin Hubble's observations that the spectral light from distant galaxies is red-shifted, indicating that the galaxies are receding from one another. In 1970, Hawking and Penrose affirmed in their Singularity Theorems that, given the assumption that the laws of General Relativity govern the universe, the universe began with an initial singularity.[18] In doing so, they removed any lingering considerations that the universe was eternal, whether this be according to the beginningless Steady-State Model or the Oscillating Model with an infinite past comprised expansions and contractions. Standard Big Bang cosmogony thus points to an absolute beginning of the universe: in the words of P. C. W. Davies, "most cosmologists think of the initial singularity as the beginning of the universe. On this view the big bang represents the creation event; the creation not only of all the matter and energy in the universe, but also of spacetime itself."[19] Thus, Hawking reports, "almost everyone now believes that the universe, and time itself, had a beginning at the big bang."[20]

Of course, modifications of the standard model have been proposed. The Hartle-Hawking "no-boundary" Quantum Gravity Model of the universe (1982) was put forth to bring together General Relativity and the world of quantum mechanics.[21] Instead of the hard-edged cone of Friedmann-LeMaître space-time model, this model proposes that, prior to Planck time, the initial stage of space-time resembled a badminton shuttlecock with a smooth, "rounded off" hemisphere without an edge. So there is no definite point of creation, even though time is still finite in the past. In his popular *Brief History of Time* (1988), Hawking would ask "What place, then, for a creator?"[22] Yet even on the Quantum Gravity model, time still began to exist, even if not at a singularity, as Hawking himself later acknowledges in his *The Grand Design* (2010).[23] In fact, the more generalized Borde-Guth-Vilenkin theorem (2003) proved that any universe which is, on average, in a state of cosmic expansion cannot be infinite in the past but must have had a beginning.[24]

The beginning of the universe—and matter, energy, space, and time—raises significant metaphysical questions regarding its cause. Again, Davies spells out what he takes to be the only options:

> "What caused the big bang?" . . . One might consider some supernatural force, some agency beyond space and time as being responsible for the big bang, or one might prefer to regard the big bang as an event without a cause. It seems to me that we don't have too much choice. Either . . . something outside of the physical world . . . or . . . an event without a cause.[25]

A second body of scientific evidence considered to support the beginning of the universe looks to the implications of its thermodynamic properties. The second law of thermodynamics states that a closed universe, with its limited amount of energy, will always move toward increased entropy. A hot object tends toward becoming a cooler object; energy inclines toward dissipation or disorder. Since the universe is winding down and, given enough time, will experience a heat death of some sort, it cannot be eternal. If the universe were eternal, why has this heat death not already occurred?

Rather than moving from a Big Bang to—as had been relatively recently assumed—a Big Crunch, in which the universe would contract and collapse back in on itself, the universe will continue to expand but move toward a cosmic Black Death, as it were. It will move toward a state of increasing coldness, disintegration, and deadness, with nothing left to show for it besides a dilute gas of fundamental particles and radiation. But if, given enough time, the universe will find itself in such a lifeless state, then it is exceedingly puzzling why, given infinite past time, the universe is not now in such a state. The most obvious solution is that the problem is based on a false assumption, namely that the universe is beginningless. Given that the universe is winding down, then in some way, as physicist Richard Schlegel acknowledges, "the universe must have been *wound up.*"[26]

So both the expansion of the universe and the second law of thermodynamics seem to point to the universe's finitude, thus offering empirical confirmation of the philosophical, deductive arguments in support of the key premise of the KCA.

This two-volume collection

As mentioned, there has been in recent decades a resurgence of interest in the KCA. A plethora of journal articles as well as entire books have been devoted to the discussion of its premises. It has now attracted the attention of physicists as well as philosophers and has found its way even into scientific journals. The time seems therefore ripe for an anthology gathering some of the disparate literature under one cover. Indeed, the breadth of the KCA literature calls for two companion volumes, one devoted to the philosophical arguments and the other to the scientific evidences. Apart from Craig and Smith's earlier co-authored *Theism, Atheism, and Big Bang Cosmology* (Clarendon Press, 1993), a brisk seller which remains in print today, there has been to date no collection or anthology devoted exclusively to a discussion of the KCA. This two-part anthology seeks to remedy that lack. It is a collection of the best recent work, both philosophical and scientific, on the premises and conclusion of the KCA.

For those who have followed the KCA, the chapters in this anthology—several of them new—will cover the familiar-sounding Hilbert's Hotel, the Tristram Shandy and Grim Reaper paradoxes, Cantorian set theory and transfinite arithmetic—as well as Big Bang cosmology, the universe's expansion, and the second law of thermodynamics. Moreover, the authors of the various chapters include both KCA advocates or sympathizers and those who contest and challenge it.

Chapter summaries of *The Kalām Cosmological Argument: Scientific Evidence for the Beginning of the Universe*

Apart from the logical conclusion of the KCA, the bulk of this second volume's content concerns the scientific evidence for the universe's beginning. Whereas the first volume introduced the first premise of the KCA as well as the philosophical arguments in support of the argument's second premise, the second volume brings us directly to the first inductive argument in support of the KCA's second premise: *The Inductive Argument Based on the Expansion of the Universe*.

Chapter 1 is excerpted from Craig's and James Sinclair's entry, "The Kalam Cosmological Argument," in the *Blackwell Companion to Natural Theology* (2009). Sinclair succinctly summarizes current cosmological theorizing with respect to the universe's beginning. He explores the implications of the Borde-Guth-Vilenkin theorem and examines the viability of a wide range of models which would avert those implications by denying the single fundamental condition upon which the theorem is based.

Chapter 2 is by J. Brian Pitts: "Why the Big Bang Singularity Does Not Help the Kalam Cosmological Argument for Theism." Pitts claims that the cosmic singularity provides "negligible" evidence for creation in the finite past, and hence theism. Physics does not require a first moment, and scientific progress will likely obviate any purported need to appeal to special divine action to explain the universe's existence.

Stimulated by Pitts's critique of the KCA and its supposed reliance on cosmogonic models featuring an initial cosmological singularity, Craig and Sinclair in Chapter 3 explore non-singular models of space-time with respect to the beginning of the universe, arguing that whether the beginning of the universe is singular or non-singular is a matter of indifference and that non-singular models have difficulty avoiding a cosmic beginning.

Chapter 4—"The Beginning of the Universe"—is written by the physicist Alexander Vilenkin. He acknowledges that the problem of the beginning of the universe has been a challenge since the early days of scientific cosmology, and various models (e.g., eternal inflation, oscillating universe) have been proffered over the years. He argues that inflation could not be eternal and must have had some sort of beginning, and an eternal universe would have reached maximum entropy, which is not what we find. The universe probably had a beginning, although theorems such as Borde-Guth-Vilenkin say nothing about God's existence. A deep mystery remains.

At this point we come to the second physical argument supporting the KCA's second premise: *The Inductive Argument Based on the Thermodynamic Properties of the Universe*. Chapter 5 ("A Dying Universe: The Long-Term Fate and Evolution of Astrophysical Objects") by Fred Adams and Gregory Laughlin discusses astrophysical issues pertaining to the long-term fate of the universe. This includes discussion of the evolution of the cosmos, the eventual exhausting of the stellar gas supply, and the destruction of the planets, stars, and galaxies. This chapter also considers "the dramatic possible effects of a non-zero vacuum energy density."

Chapter 6 is Gábor Kutrovátz's essay, "Heat Death in Ancient and Modern Thermodynamics." Heat death, he notes, is useful for understanding the behavior and evolutionary features of various thermodynamic systems. This piece discusses Aristotle's physics, which purportedly includes an implicit solution to the problem of heat death and thus offers a fruitful comparison with the different solution presented by modern thermodynamics.

Chapter 7—Milan M. Cirkovic's "Entropy and Eschatology: a Comment on Kutrovátz's Paper 'Heat Death in Ancient and Modern Thermodynamics'"—challenges Kutrovátz's claims and attempts to point out alleged errors he makes about cosmology. Cirkovic states that another pathway—an ancient philosophical tradition—affords answers to the heat death puzzle. This pathway includes the "final duration of time"—confirmed by recent rediscovery—as well as the notion of "observational self-selection," as noted in recent articulations of the anthropic principle. These authors claim that understanding the future evolution and final state of the universe ("physical eschatology") can be achieved without appealing to metaphysics.

Chapter 8, "The Generalized Second Law Implies a Quantum Singularity Theorem" by Aron C. Wall, asserts that the second law of thermodynamics can help prove a singularity theorem: because space is finite, only a finite amount of entropy-producing processes must have existed in the past *unless* there could be some reversal of the arrow of time, which he thinks "would seem to raise the same sorts of philosophical questions that any other sort of beginning in time would."[27]

This then brings us to the conclusion of the KCA: *Therefore, the universe has a cause.* This section includes three chapters. The first of these is J. P. Moreland's "Libertarian Agency and the Craig/Grünbaum Debate about Theistic Explanation of the Initial Singularity" (Chapter 9). Craig and Grünbaum have debated the KCA's conclusion, which points to a personal cause for the beginning of the universe. Moreland brings recent discussion on libertarian agency theory to bear on this debate, which he sees as strengthening Craig's case against Grünbaum. The resources of libertarian agency theory, he insists, can overcome and rebut Grünbaum's five main objections to theistic personal explanation.

Chapter 10, "Causation and the Logical Impossibility of a Divine Cause" by Quentin Smith, acknowledges that while virtually all contemporary theists, agnostics, and atheists believe that a divine cause of the universe is logically possible, the root of the problem is that philosophers have failed to undertake the kind of necessary investigation into the nature of causation. All extant definitions of causality logically rule out an originating divine cause.

The final chapter of this second volume of the anthology, "Beyond the Big Bang," is an excerpt of an article "Naturalism and Cosmology" by Craig, in which he responds to Quentin Smith's essay preceding his. He argues that Smith's assertion that God's causing the universe is logically inconsistent with all extant definitions of causality is not merely unjustified and false but, far more important, of little significance, even if true. Craig concludes that it is far more plausible to deny at least one of Smith's premises than to affirm what Hume called the 'absurd Proposition' that something might arise without a cause.

It is my hope that this and its companion volume will serve well those who are upper-level philosophy students and professors engaging in philosophy of religion, philosophical theology, and the dialogue between science and theology. But those involved in the following disciplines and areas of study will also profit from engaging in this conversation: Cosmogony, Astrophysics, Physics (time, space, causality, thermodynamics), Philosophy of science, Mathematics (Cantorian set theory, infinity), Metaphysics (abstract objects, agent causation), and Theology (the doctrine of God, creation, divine agency).

Notes

1 London: Macmillan, 1979.
2 "Kalam Cosmological Arguments for Atheism," in *The Cambridge Companion to Atheism*, ed. Michael Martin (Cambridge: Cambridge University Press, 2007), 183.
3 As further indication of a strong interest in the KCA, an anthology of essays on the KCA is being published in Arabic by Jamalon Publishers (2017). So far as I can tell, several of the same outstanding articles that I have selected will be translated into Arabic for that volume.
4 A beloved professor who became a dear friend to both Craig and me, Hackett was a brilliant, eccentric, amiable teacher and a very capable philosopher noted for his paragraph-long, Germanic-style sentences. He had been Craig's professor at Wheaton College, Illinois, and Craig would later become Hackett's colleague at Trinity Evangelical Divinity School. I had the privilege of studying under both these scholars, writing my M.A. thesis at Trinity on the KCA under the supervision of Hackett as my first reader.
5 Originally published by Moody Press of Chicago, the book has since been republished by Wipf and Stock (2009).
6 Hackett's follow-up defense of the Christian revelation claim also propounds this argument: *The Reconstruction of the Christian Revelation Claim: A Philosophical and Critical Apologetic* (Eugene, OR: Wipf and Stock, 2009), 98–103. This book was previously published by Baker Book House in 1984.
7 Stuart C. Hackett, *The Resurrection of Theism* (Eugene, OR: Wipf and Stock, 2009), 293–94.
8 *The Cosmological Argument from Plato to Leibniz* (London: Macmillan, 1979).
9 Thomas Aquinas, *Summa Theologica* in *Introduction to St. Thomas Aquinas*, ed. Anton Pegis (New York: Modern Library, 1948).
10 G. W. Leibniz, "The Principles of Nature and of Grace, Based on Reason," in *Leibniz Selections,* ed. P. Wiener (New York: Charles Scribner's Sons, 1951), 527.
11 G. W. Leibniz, "The Monadology," in *Leibniz Selections,* 539.
12 Bertrand Russell and Frederick Copleston, "The Existence of God," *The Existence of God*, ed. John Hick (New York: Macmillan, 1964), 175.
13 *Is Faith in God Reasonable?* With Alex Rosenberg. Ed. Paul Gould and Corey Miller, R*outledge Studies in the Philosophy of Religion* (London: Routledge, 2014).
14 Al-Ghazali, *Kitab al-Iqtisad fi'l-Iqtiqad* (Ankara: University of Ankara Press, 1962), 15–16.

15 *God Over All: Divine Aseity and the Challenge of Platonism* (Oxford: Oxford University Press, 2016); *God and Abstract Objects: The Coherence of Theism III: Aseity* (Berlin: Springer Verlag, 2017).
16 *The Tensed Theory of Time: A Critical Examination*, Synthese Library 293 (Dordrecht: Kluwer Academic Publishers, 2000); *The Tenseless Theory of Time: A Critical Examination*, Synthese Library 294 (Dordrecht: Kluwer Academic Publishers, 2000).
17 A. Friedmann, "Über die Krümmung des Raumes," *Zeitschrift für Physik* 10 (1922): 377–86; G. Lemaître, "Un univers homogène de masse constante et de rayon croissant, rendant compte de la vitesse radiale des nébuleuses extragalactiques," *Annales de la Société scientifique de Bruxelles* 47 (1927): 49–59.
18 R. Penrose, "Gravitational Collapse and Space-Time Singularities," *Physical Review Letters* 14 (1965): 57–59; S. W. Hawking and R. Penrose, *The Large-Scale Structure of Space-Time,* ed. S. W. Hawking and G. F. R. Ellis (Cambridge: Cambridge University Press, 1973), 266.
19 P. C. W. Davies, "Spacetime Singularities in Cosmology," in *The Study of Time III,* ed. J. T. Fraser (New York: Springer Verlag, 1978), 78–79.
20 Stephen Hawking and Roger Penrose, *The Nature of Space and Time,* The Isaac Newton Institute Series of Lectures (Princeton, NJ: Princeton University Press, 1996), 20.
21 S. Hawking and J. Hartle, "The wave function of the universe," *Physical Review D* 28 (1983): 12, 2960–75.
22 Stephen Hawking, *A Brief History of Time* (New York: Bantam, 1988), 141.
23 Leonard Mlodinow and Stephen Hawking, *The Grand Design* (New York: Bantam, 2010).
24 A. Borde, A. Guth, and A. Vilenkin (2003): "Inflationary Spacetimes are not Past-complete," *Physical Review Letters* (2003): 90, 151301, preprint: http://arxiv.org/abs/gr-qc/0110012.
25 Paul Davies, "The Birth of the Cosmos," in *God, Cosmos, Nature and Creativity*, ed. Jill Gready (Edinburgh: Scottish Academic Press, 1995), 8–9.
26 Richard Schlegel, "Time and Thermodynamics," in *The Voices of Time*, ed. J. T. Fraser (London: Penguin, 1948), 511.
27 Aron C. Wall, "The Generalized Second Law implies a Quantum Singularity Theorem," arXiv: 1010.5513v3 [gr-qc] January 24, 2013, p. 38.

1

The *Kalām* Cosmological Argument: "Science" Excerpt

William Lane Craig and James D. Sinclair

2.3 Scientific confirmation

The sort of philosophical problems with the infinity of the past which have been the object of our discussion are now being recognized in scientific papers by leading cosmologists and philosophers of science.[1] For example, Ellis, Kirchner, and Stoeger ask, "Can there be an infinite set of really existing universes? We suggest that, on the basis of well-known *philosophical* arguments, the answer is No" (Ellis *et al.* 2003, p. 14, our emphasis). Similarly, noting that an actual infinite is not constructible and therefore not actualizable, they assert, "This is precisely why a realized past infinity in time is not considered possible from this standpoint—since it involves an infinite set of completed events or moments" (Ellis *et al.* 2003, p. 14). These misgivings represent endorsements of both the *kalām* arguments defended above. Ellis and his colleagues conclude, "The arguments against an infinite past time are strong—it's simply not constructible in terms of events or instants of time, besides being conceptually indefinite" (Ellis *et al.* 2003, p. 14).

Apart from these philosophical arguments, there has emerged during the course of the twentieth century provocative empirical evidence that the universe is not past eternal. This physical evidence for the beginning of the universe comes from what is undoubtedly one of the most exciting and rapidly developing fields of science today: astronomy and astrophysics. Prior to the 1920s, scientists had always assumed that the universe was stationary and eternal. Tremors of the impending earthquake that would topple this traditional cosmology were first felt in 1917, when Albert Einstein made a cosmological application of his newly discovered gravitational theory, the General Theory of Relativity (Einstein 1917, pp. 177–88). In so doing he assumed that the universe is homogeneous and isotropic and that it exists in a steady state, with a constant mean mass density and a constant curvature of space. To his chagrin, however, he found that General Relativity (GR) would not permit such a model of the universe unless he introduced into his gravitational field equations a certain "fudge factor" Λ in order to counterbalance the gravitational effect of matter and so ensure a static universe. Einstein's universe was balanced on a razor's edge, however, and

the least perturbation—even the transport of matter from one part of the universe to another—would cause the universe either to implode or to expand. By taking this feature of Einstein's model seriously, the Russian mathematician Alexander Friedmann and the Belgian astronomer Georges Lemaître were able to formulate independently in the 1920s solutions to the field equations which predicted an expanding universe (Friedmann 1922).

Friedmann's first equation is:

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho + \frac{\Lambda}{3} - \frac{k}{a^2}$$

where H = Hubble parameter, a = scale factor, G = gravitational constant, ρ = mass density of universe, Λ = cosmological constant, K = curvature parameter, c = speed of light. By way of explanation, the scale factor, "a," of the universe is a global multiplier to universe size. Imagine the universe as a perfectly laid out city with streets that travel only north-south and east-west. Streets are spaced at equal distance intervals. Street intersections then define perfectly symmetric city blocks. One could go further and think of buildings in the city as analogous with galaxies in the universe.

The distance from one city block to another is a function of two values: the originally laid out distance (call that the "normalized" distance) and the scale factor multiplier "a." Note that, as in Fig. 1.1, when one multiplies by a scale factor of 1/2, one still has precisely the same city with the same number of city blocks. The only thing that has changed is the distance interval between the city blocks.

Now consider buildings within the city. If the city block distance were reduced to the size of buildings, clearly something must give. The buildings would be squeezed together and destroyed. This is analogous to what happens with matter in the real universe. The sizes of non-elementary mass structures such as protons, neutrons, atomic nuclei, and so on, are fixed; they do not change with the scale factor. Other physical structures, such as massless particles, *do* adjust with the scale factor. The wavelength of

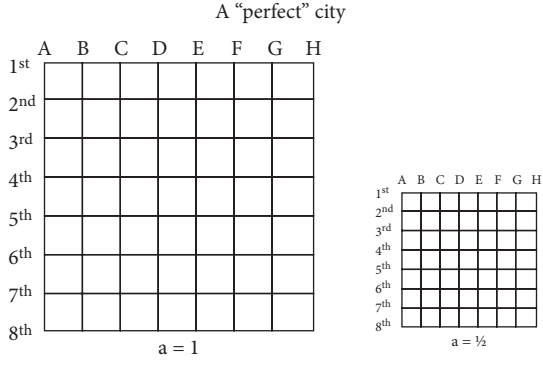

Figure 1.1 Analogy of the universe as a city laid out in a grid.

radiation adjusts and hence would gain (in the case of contraction) or lose (in the case of expansion) energy as a result.[2] When one crowds these particles in upon themselves, one will see a transition to different physics.

Recall that the full city is always present regardless of the value of the scale factor. So now consider two additional situations. First, suppose that the scale factor were to shrink to zero. Space (and time) would disappear. Any structure that could not transform to zero size would be destroyed. If there were no physical process that would allow such a thing to happen, we should seem to have a paradox. Either there must be an undiscovered physical process or the scale factor cannot, in reality, assume a null value.

Second, imagine that the city is of infinite size. Conceptually, there is no problem with extending the streets north-south and east-west to infinity in each direction. What does it mean to scale the universe's size in such a situation? No matter what scale factor one adopts, the size of the universe remains infinite. Nevertheless, the idea of scaling still retains coherency in that we can apply a multiplier to the finite distance between city blocks. Yet what would be the meaning of applying a zero scale factor in this situation? Now it would appear that the size of the full universe is "zero times infinity" which, in general, can be any finite number (Barrow 2005, p. 160). What does this mean, given that the distance between *any* spot in the universe to *any* other spot must still be zero? GR simply breaks down at zero scale factor.

Whether or not the full universe is of infinite or finite size is given in the Friedmann equation by the curvature parameter "K." A positive K indicates that the universe, much like the surface of the Earth, is unbounded yet of finite size. Going back to the analogy, imagine that the city is laid out over Earth's entire curved surface. A traveler on 1st street would never come to the end; rather he would eventually come back to the location where he started. A positive K yields positive curvature and a closed universe. This is one type of "compact metric" within GR.

A zero value for "K" yields a "flat" universe. 1st street is unbounded and of infinite length (in both directions). A similar situation obtains for a negative K value. Here one has negative, or "saddle-shaped" curvature. Two travelers moving east and side-by-side up 1st and 2nd streets would actually get laterally farther apart as the curvature of the surface causes the streets to diverge from each other. The latter case gives an infinitely sized "open" universe.

The components of the universe (all the energy, keeping in mind that $E = mc^2$) determine what type of curvature the universe possesses. The "strength" of gravity, included in the equation via the parameter "G," affects the magnitude of the curvature.

The parameters ρ and Λ indicate the type and magnitude of the different types of energy that cause the curvature. The parameter "ρ" represents the density (that is, the energy per unit volume) of the two types of "ordinary" energy: matter and radiation. It is "ordinary" in the sense that we are familiar with it in daily life and it is of a form that makes gravity an attractive force. Λ represents an exotic type of energy density which can transform gravity from an attractive to a repulsive force.

Friedmann's first equation tells us how the scale factor changes as time elapses. Mathematically, this is the first derivative of the scale factor "a," known as "a-dot," or

å. One can see that the increase (or decrease) in the scale factor is strongly a function of the universe's energy content. Now the "ordinary" energy density ρ will become smaller as the universe expands, since one has the same amount of energy spread out over a greater volume. So its causal impact on the expansion will progressively diminish at ever later times (this works in reverse for contraction). By contrast, Λ, which represents the dark energy density, is constant. The dark energy does not become more dilute during expansion or concentrated during contraction. Hence, early in the life of an expanding universe Λ is unimportant compared to ρ. But its impact "snowballs" as time goes on. As long as the impact of ρ in the early universe is not enough to overturn an expansion and begin a contraction, the effect of Λ will eventually lead to a runaway expansion of the universe. There will appear a moment in the history of the universe when the dark energy will begin to dominate the ordinary energy, and the universe's expansion will begin to accelerate. Recent observations, in fact, seem to show precisely this effect in our own universe, with a transition age at 9 billion years (Overbye 2006).

Friedmann's second equation gives the rate of change of the expansion rate:

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3\mathrm{p}) + \frac{\Lambda}{3}$$

It determines whether the expansion itself is slowing or accelerating. This acceleration is referred to as "a-double-dot," or ä. A new term "*p*" appears in the equation. This is the pressure (similar to the pressure of a gas inside a balloon). Pressure itself can produce gravitational force. Pressure is normally negligible for ordinary matter, although it can play a role in radiation dominated universes. Pressure, however, can have a tremendous impact given a universe dominated by dark energy. As Friedmann's second equation shows, the rate at which the expansion of the universe accelerates is proportional to: (- μ - 3*p*). But the pressure in the vacuum is just equal to the negative of the energy density; this is called the equation of state. Hence, overall, the acceleration is positive (which will produce expansion) and proportional to twice the energy density.

Ordinary matter will exert positive pressure (which will keep a balloon inflated, for example). This type of pressure will produce an attractive gravitational force, which supplements the attractive gravity that accrues from mass. Dark energy has the bizarre property that it generates *negative* pressure. But dark energy, while it has a positive energy density (which contributes to attractive gravity) will, *on net*, produce a repulsive gravitational effect. Looking at Friedmann's second equation, one sees that an attractive gravitational contribution tends to slow down expansion (or accelerate contraction), while repulsive gravity will do the opposite.

The monumental significance of the Friedmann-Lemaître model lay in its historization of the universe. As one commentator has remarked, up to this time the idea of the expansion of the universe "was absolutely beyond comprehension. Throughout all of human history the universe was regarded as fixed and immutable and the idea that it might actually be changing was inconceivable" (Naber 1988,

pp. 126–27). But if the Friedmann-Lemaître model is correct, the universe can no longer be adequately treated as a static entity existing, in effect, timelessly. Rather the universe has a history, and time will not be a matter of indifference for our investigation of the cosmos.

In 1929 the American astronomer Edwin Hubble showed that the red-shift in the optical spectra of light from distant galaxies was a common feature of all measured galaxies and was proportional to their distance from us (Hubble 1929, pp. 168–73). This red-shift, first observed by Vesto Slipher at the Lowell Observatory,[3] was taken to be a Doppler effect indicative of the recessional motion of the light source in the line of sight. Incredibly, what Hubble had discovered was the isotropic expansion of the universe predicted by Friedmann and Lemaître on the basis of Einstein's GR. It was a veritable turning point in the history of science. "Of all the great predictions that science has ever made over the centuries," exclaims John Wheeler, "was there ever one greater than this, to predict, and predict correctly, and predict against all expectation a phenomenon so fantastic as the expansion of the universe?" (Wheeler 1980, p. 354).

2.31 The standard hot Big Bang model

According to the Friedmann-Lemaître model, as time proceeds, the distances separating the ideal particles of the cosmological fluid constituted by the matter and energy of the universe become greater. It is important to appreciate that as a GR-based theory, the model does not describe the expansion of the material content of the universe into a pre-existing, empty, Newtonian space, but rather the expansion of space itself. The ideal particles of the cosmological fluid are conceived to be at rest with respect to space but to recede progressively from one another as space itself expands or stretches, just as buttons glued to the surface of a balloon will recede from one another as the balloon inflates. As the universe expands, its density progressively declines.

This has the astonishing implication that as one reverses the expansion and extrapolates back in time, the universe becomes progressively denser until one arrives at a state of infinite density at some point in the finite past. This state represents a singularity at which space-time curvature, along with temperature, pressure, and density, becomes infinite. To be more correct, the volume of the universe *approaches* zero in the limit as the *scale factor* of the universe approaches zero. The Friedmann-Lemaître model does not, in fact, describe what happens at the singularity, since Einstein's GR breaks down at this limit.

The initial cosmological singularity is therefore not in space-time but constitutes an edge or boundary to space-time itself. Robert Wald describes how singular spacetimes are to be properly characterized:

> By far the most satisfactory idea proposed thus far is basically to use the "holes" left behind by the removal of singularities as the criterion for their presence. These "holes" should be detectable by the fact that there will be geodesics which have finite affine length; that is, more precisely there should exist geodesics which

> are inextendible in at least one direction but have only a finite range of affine parameter. Such geodesics are said to be *incomplete*. (For timelike and spacelike geodesics, finite affine "length" is equivalent to finite proper time or length so the use of affine parameter simply generalizes the notion of "finite length" to null geodesics.) Thus, we could define a spacetime to be singular if it possesses at least one incomplete geodesic.
>
> Nevertheless, there is a serious physical pathology in any spacetime which is timelike or null geodesically incomplete. In such a spacetime, it is possible for at least one freely falling particle or photon to end its existence within a finite "time" (that is, affine parameter) or to have begun its existence a finite time ago. Thus, even if one does not have a completely satisfactory general notion of singularities, one would be justified in calling such spacetimes physically singular. It is this property that is proven by the singularity theorems to hold in a wide class of spacetimes (Wald 1984, pp. 215–16).[4]

The existence of a boundary to space-time implies, not merely that the "stuff" of the universe begins to exist, but that space and time do as well (for in the Friedmann-Lemaître model all past-directed geodesics terminate at the singularity). P. C. W. Davies comments,

> If we extrapolate this prediction to its extreme, we reach a point when all distances in the universe have shrunk to zero. An initial cosmological singularity therefore forms a past temporal extremity to the universe. We cannot continue physical reasoning, or even the concept of spacetime, through such an extremity. For this reason most cosmologists think of the initial singularity as the beginning of the universe. On this view the big bang represents the creation event; the creation not only of all the matter and energy in the universe, but also of spacetime itself (Davies 1978, pp. 78–9).

The term "Big Bang," originally a derisive expression coined by Fred Hoyle to characterize the beginning of the universe predicted by the Friedmann-Lemaître model, is thus potentially misleading, since the expansion cannot be visualized from the outside (there being no "outside," just as there is no "before" with respect to the Big Bang).[5]

The standard hot Big Bang model, as the Friedmann-Lemaître model came to be called, thus describes a universe which is not eternal in the past, but which came into being a finite time ago. Moreover,—and this deserves underscoring—the origin it posits is an absolute origin *ex nihilo*. For not only all matter and energy, but space and time themselves come into being at the initial cosmological singularity. As Barrow and Tipler emphasize, "At this singularity, space and time came into existence; literally nothing existed before the singularity, so, if the Universe originated at such a singularity, we would truly have a creation *ex nihilo*" (Barrow and Tipler, 1986, p. 442). On such a model the universe originates *ex nihilo* in the sense that it is false that something existed prior to the singularity.

2.32 Evidence for general relativity

The earliest evidence in favor of the Big Bang came from the consonance of theory and experiment. Einstein's early papers proposed two tests that could be performed immediately. It had been known for some time that Newton's gravitational theory could not adequately describe the orbit of the planet Mercury. The real orbit precessed around the sun (that is, the ellipse itself rotates over time). In contrast to Newton's theory of gravity, GR "predicted" that this precession should take place. Einstein's theory also predicted that, since matter bends space, light rays should have their paths noticeably bent when they pass close to massive objects. A solar eclipse in 1919 provided the opportunity for a test of this prediction. An expedition lead by Arthur Eddington confirmed that light rays were indeed deflected.

These tests were not sufficiently accurate[6] to ensure that small deviations from GR were not possible. It was also suspected that, since the real universe is not completely homogeneous and isotropic at all scales, Friedmann and Lemaître's prediction of a true singular beginning to the universe would ultimately fail. Perhaps a slight anisotropy could result in matter's "sling-shotting" past itself at a minimum (but non-zero) radius condition, so that the present expansion was preceded by a cosmic contraction, thereby avoiding the absolute beginning of the universe. In 1970, however, Stephen Hawking and Roger Penrose proved that the homogeneity/isotropy assumption was irrelevant. The Hawking-Penrose singularity theorems showed that so long as the universe is governed by GR (with a few technical exceptions which will become prominent in our discussion below), our past must include a singularity (Hawking and Penrose 1970). Wald comments:

> [Hawking-Penrose 1970] gives us strong reason to believe that our universe is singular . . . the observational evidence strongly suggests that our universe – or, at least, the portion of our universe within our causal past – is well described by a Robertson-Walker model [standard hot Big Bang theory] at least back as far as the decoupling time of matter and radiation. However, in these models, the expansion of the past directed null geodesics emanating from the event representing us at the present time becomes negative at a much more recent time than the decoupling time. Thus there is strong reason to believe that condition 4c of [Hawking-Penrose 1970] is satisfied in our universe. Since we expect that conditions (1)-(3) also are satisfied, it appears that our universe must be singular. Thus, it appears that we must confront the breakdown of classical general relativity expected to occur near singularities if we are to understand the origin of our universe (Wald 1984, p. 241).

The conditions Wald mentions are:

1. Satisfaction of the strong energy condition (typically obeyed by "normal" types of matter).
2. Satisfaction of the generic energy condition (there is no exotic property of the spacetime that prevents gravitational focusing).

3. No closed time loops (the future does not bend back and become one's own past).
4. There is a point p such that past-directed worldlines emanating from p have negative expansion; that is, they are focused back on each other (the worldlines of observers do not trace into an infinite past given sufficiently concentrated matter but are "focused," as by a lens, into a singular condition within a finite time).

2.33 Exceptions to the Hawking-Penrose theorems

Four possible exceptions to the Hawking-Penrose singularity theorems conveniently distinguish four classes of non-standard models which provide possible alternatives to the standard Big Bang model (Fig. 1.2). The H-P theorem also has the obvious, but implicit, condition that GR is fundamental; that is, it is a complete as well as correct description of conditions within our universe (thereby defining a 5th condition).

The first option (closed time loops) has been the subject of some exploration in cosmological circles. The next two—eternal inflation and quantum gravity—represent areas of fertile cosmological investigation which merit our attention. The last two exception conditions are not expected to be part of "reasonable" physical models of the universe. Hawking explains:

> Between 1965 and 1970 Penrose and I used the techniques I have described to prove a number of singularity theorems. These theorems had three kinds of conditions. First there was an energy condition such as the weak, strong or generic energy conditions. Then there was some global condition on the causal structure such as that there shouldn't be any closed time like curves. And finally was some condition that gravity was so strong in some region that nothing could escape. . . .

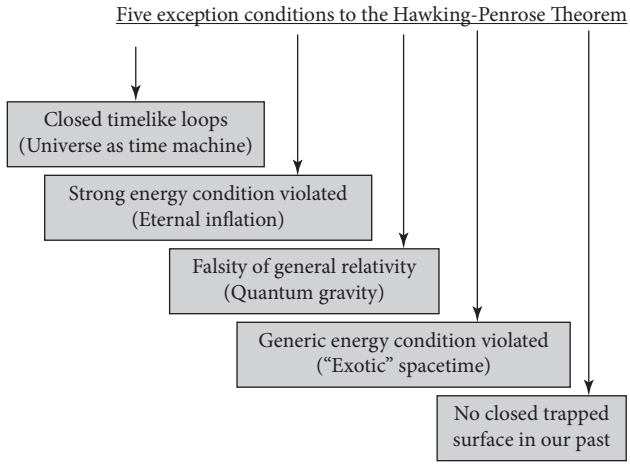

Figure 1.2 Model classes based on exceptions to the H-P singularity theorems.

> [The generic energy condition] says that first the strong energy condition holds. Second, every timelike or null geodesic encounters some point where there is some curvature that is not specially aligned with the geodesic. The generic energy condition is not satisfied by a number of known exact solutions. But these are rather special. One would expect it to be satisfied by a solution that was "generic" in an appropriate sense [that is, a reasonable physical model]. If the generic energy condition holds, each geodesic will encounter a region of gravitational focusing (Hawking and Penrose 1996, pp. 14–15).

We do expect our past to feature a closed, trapped surface, and there is no reason to postulate an exotic spacetime construction that would just happen to have perfect defocusing characteristics so as to counter the effects of gravity. Hence, our discussion will revolve around the first three options.

I Closed timelike curves

A first, exotic exception to the Hawking-Penrose theorems is the possible existence of closed, timelike curves. Permitted by Einstein's GR, closed, timelike curves represent an observer tracing out a circular path through space and time.

J. Richard Gott and Li-Xin Li have proposed a model according to which the early universe (only) is a closed time loop that occasionally gives "birth" to a universe like ours (Fig. 1.3). They maintain that Alexander Vilenkin's "tunneling from nothing" model (see section IVc) should properly be taken as tunneling from a previous state. As will be seen in this essay, most cosmological models assert that the past terminates at a boundary a finite time ago. One then wishes to explain what exists at that boundary. Vilenkin (and, independently, the team of Stephen Hawking and James Hartle) believes that the universe sprang into being "out of nothing." Gott and Li believe, instead, that there is a closed timelike curve (CTC) at this boundary.

The Gott-Li model could be an example of a universe timeline looping back upon itself to allow the universe, in Gott and Li's words, "to become its own mother." They explain:

> In this paper, we consider instead the notion that the Universe did not arise out of nothing, but rather created itself. One of the remarkable properties of the theory of general relativity is that in principle it allows solutions with CTCs. Why not apply this to the problem of the first-cause?
>
> Many inflationary models allow creation of baby inflationary universes inside black holes, either by tunneling across the Einstein-Rosen bridge, or by formation as one approaches the singularity. *If one of these baby universes simply turns out to be the universe we started with, then a multiply connected model with early CTCs bounded by a Cauchy horizon is produced.*
>
> . . . Then the Universe neither tunneled from nothing, nor arose from a singularity; it created itself (Gott and Li 1998, p. 39, emphasis ours).

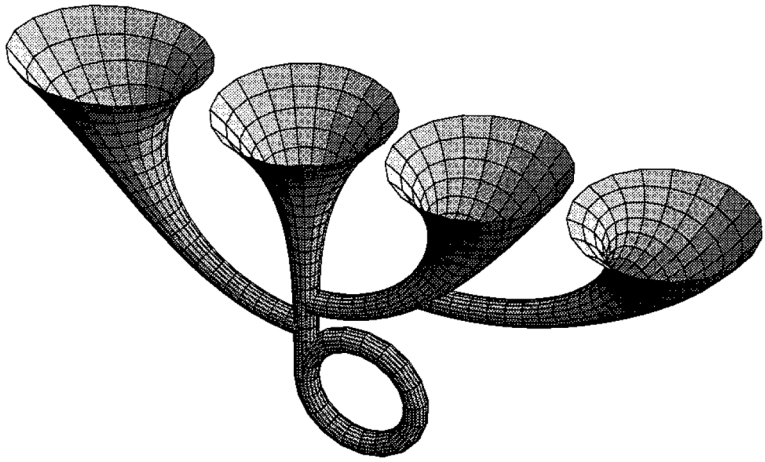

Figure 1.3 A Gott-Li Universe Time Machine: The region of closed timelike curves exists at the bottom of the diagram. This region is separated from the future evolution of the universe by a Cauchy horizon [a boundary that separates spacetime into CTC regions and non-CTC regions]. The four branches to the top of the diagram can be thought of as inflationary bubbles undergoing a de Sitter-like expansion [see section II & IIIa for a discussion of inflation and de Sitter spaces].

Some histories in our past were circular in nature (forming a multiply connected spacetime) and recreating the Big Bang. Further, this is not a cyclic universe; *it is the same Big Bang*.

The CTC scenario raises interesting philosophical questions about the nature of time (see p. 191 in original Craig and Sinclair essay). But here our interest is in the model's physical viability. The primary physical problem confronting CTC models in general is their violation of the so-called Chronology Protection Conjecture.

Gott and Li indicate that "the region of CTCs . . . should be in a pure vacuum state containing no real particles or Hawking radiation and no bubbles" (Gott and Li 1998, p. 39). This is so because this stray radiation would destroy the CTC. The reason for this curious feature of a CTC model was discussed by Stephen Hawking in (Hawking 1992), where he formally suggested a "Chronology Protection Conjecture." His theory was that a time machine (CTC) would have characteristics that were so unstable that it would quickly destroy itself. Hence nature conspires to prevent time machines. A popular level (and entertaining) description of this effect is given by GR theorist Kip Thorne. He constructs a scenario that allows a local time machine to exist with one end on a spaceship departing Earth with his wife Carole and the other end on Earth with him in his living room (this is an exotic general relativistic structure called a wormhole).

> Imagine that Carole is zooming back to Earth with one wormhole mouth in her spacecraft, and I am sitting at home on Earth with the other. When the spacecraft

> gets to within 10 light-years of Earth, it suddenly becomes possible for radiation (electromagnetic waves) to use the wormhole for time travel: any random bit of radiation that leaves our home in Pasadena traveling at the speed of light toward the spacecraft can arrive at the spacecraft after 10 years' time (as seen on Earth), enter the wormhole mouth there, travel back in time by 10 years (as seen on Earth), and emerge from the mouth on Earth at precisely the same moment as it started its trip. The radiation piles right on top of its previous self, not just in space but in spacetime, doubling its strength. What's more, during the trip each quantum of radiation (each photon) got boosted in energy due to the relative motion of the wormhole mouths (a "Doppler-shift" boost).
>
> After the radiation's next trip out to the spacecraft then back through the wormhole, it again returns at the same time as it left and again piles up on itself, again with a Doppler-boosted energy. Again and again this happens, making the beam of radiation infinitely strong.
>
> In this way, beginning with an arbitrarily tiny amount of radiation, a beam of infinite energy is created, coursing through space between the two wormhole mouths. As the beam passes through the wormhole . . . it will produce infinite spacetime curvature [*i.e.*, a singularity] and probably destroy the wormhole, thereby preventing [a time machine from coming into being in the first place] (Thorne 1994, pp. 505–06).

Of interest to us is the general applicability of this effect to the Gott-Li model. Gott and Li are sensitive to this problem and have developed a solution.[7] They and others have found some specially constructed spacetimes that appear to elude Hawking's Chronology Protection Conjecture (CPC). To avoid the CPC, they have constructed a special initial state for the universe: a zero temperature empty space called an "adapted Rindler vacuum." It is specially built and balanced such that it does not develop the destructive effect suggested by Thorne above.

After the publication of Gott and Li's paper, William Hiscock developed a defense of the CPC that still appears to stand (Hiscock 2000). First, Hiscock argues that the Gott-Li choice of initial condition is highly fine-tuned. In fact, Gott-Li's vacuum is of "measure zero" in the set of all possible Rindler vacuums. This means that the scenario is just about as unlikely as is possible without ruling it out summarily. D. H. Coule agrees in his summary of quantum gravity models, referring to the Gott-Li model as "rather contrived" (Coule 2005).[8] Second, Hiscock argues that the Gott-Li vacuum is not stable, given more realistic physical force fields. He writes:

> . . . the (Rindler) vacuum stress-energy of a nonconformally coupled scalar field, or a conformally coupled massless field with a . . . self-interaction will diverge on the chronology horizon for all values of the Misner identification scale [this is the parameter that Gott-Li have fine-tuned]. In addition, the vacuum polarization of [the scalar field considered in the Gott-Li model] diverges in all cases [leading to the Thorne effect cited above], even for the conformally invariant case examined by Li and Gott. Hence, the regular behavior found by Cassidy and Li and Gott

> holds only for a conformally invariant, non-interacting field, and only for the stress-energy tensor. While some fields in nature (*e.g.*, the electromagnetic field, before interactions are added) are conformally invariant, others – notably gravity itself–are not; and interactions are the rule, not the exception (Hiscock 2000, p. 4).

Coule adds: ". . . in Misner space this state [Gott-Li model] was only possible with identification scale b = 2π, or b = 2πr0 for the multiple de Sitter case. Such an exact value is itself inconsistent with notions of quantum uncertainty" (Coule 2005, p. 31). So the Heisenberg uncertainty principle of quantum mechanics would guarantee that the relevant parameter could not be "just-so." But if it isn't "just-so," then the universe collapses into a singular condition in the presence of a time machine. Coule also suggests that this parameter, called the "Misner identification scale" is not a constant. Rather it is likely to change dynamically as a function of matter couplings or energy potentials. As soon as it does, the CTC will destabilize.

Interestingly, Gott and Li used similar objections when arguing for their model at the expense of the "creation from nothing" approach of Vilenkin and Hartle-Hawking [section IVc]. Gott-Li criticize the "creation from nothing" approach on the grounds of the uncertainty principle and the fact that their competitors are not using realistic force fields; that is to say, the Vilenkin approach is not close enough to what we expect for the real universe. Yet their own model appears to break down when similar objections are leveled against *it*.[9, 10]

CTC physics is interesting, and while some theorists still pursue it, it occupies only a small minority of ongoing cosmological investigation. While it is true that no one has been able definitively to rule out CTCs, the evidentiary burden lies upon those defending the viability of such spacetimes and models predicated upon their reality.

II. Eternal inflation

Motivation

A more serious exception to the Hawking-Penrose singularity theorems is afforded by inflationary theory. Although the Friedmann-Lemaître model had a great deal of evidential support, there were, nonetheless, observational anomalies which suggested that there was more to the story. There were also theoretical reasons to think that the description was not quite complete. These difficulties, especially the horizon, flatness, and cosmic relic problems, prompted theorists to propose a modification of the standard Big Bang picture called "inflation."

With respect to the horizon problem, cosmologists lacked an explanation as to *why* the universe should be so homogeneous and isotropic.[11] Alan Guth explains,

> The initial universe is assumed to be homogeneous, yet it consists of at least $\sim 10^{83}$ separate regions which are causally disconnected (*i.e.*, these regions have not yet had time to communicate with each other via light signals) . . . Thus, one must assume that the forces which created these initial conditions were capable of violating causality (Guth 1981, p. 347).

Cosmology had an appropriate "organizing" principle at hand – thermodynamic equilibrium – , yet mathematics showed that, in the limit, as one looked backwards at the Big Bang, the different parts of the universe would lose causal communication. Without causal communication, all the parts of the observable universe could not have cooperated in energy transfer so as to make all parts look the same (in the present). Physicist Brian Greene describes the horizon problem:

> Physicists define a region's *cosmic horizon* (or *horizon* for short) as the most distant surrounding regions of space that are close enough to the given region for the two to have exchanged light signals in the time since the [Big] bang. . . . The *horizon problem*, then is the puzzle, inherent in the observations, that regions whose horizons have always been separate – regions that could never have interacted, communicated, or exerted any kind of influence on each other – somehow have nearly identical temperatures (Greene 2004, p. 289, emphasis in original).
>
> . . . imagine running the cosmic film in reverse while focusing on two regions of space currently on opposite sides of the observable universe – regions that are so distant that they are beyond each other's spheres of influence. If in order to halve their separation we have to roll the cosmic film more than halfway back toward the beginning, then even though the regions of space were closer together, communication between them was still impossible: they were half as far apart, but the time since the bang was *less* than half of what it is today, so light could travel only *less* than half as far. Similarly, if from that point in the film we have to run more than halfway back to the beginning in order to halve the separation between the regions once again, communication becomes more difficult still. With this kind of cosmic evolution, even though regions were closer together in the past, it becomes more puzzling – not less –that they somehow managed to equalize their temperatures. Relative to how far light can travel, the regions become increasingly cut off as we examine them ever farther back in time. This is exactly what happens in the standard big bang theory (Greene 2004, p. 288, emphasis in original).

A second problem was that the universe appears to be "flat" (that is, space is Euclidian: the angles of a triangle add up to 180 degrees; parallel lines do not intersect), while GR predicts that that is a wildly improbable outcome.

> A typical closed universe will reach its maximum size on the order [of the Planck scale of 10^{-44} sec], while a typical open universe will dwindle to a ρ [density] much less than ρ_{cr} [critical density; the density for a long-lived universe]. A universe can only survive ~ 10^{10} years [approximately the age of our universe] only by extreme fine tuning. . . . For [the likely initial conditions for our universe] the value of H_0 [the initial expansion rate of the universe] must be fine tuned to an accuracy of one part in 10^{55}. In the standard model this incredibly precise initial relationship must be assumed without explanation (Guth 1981, p. 348).

The third problem was that the supposition of an initial disorganized state of the universe led to the prediction of the presence of bizarre cosmic relics. Magnetic

monopoles should appear in our universe at a density amenable to detection with our present means. In Guth's original paper on inflation, he indicates that standard particle physics predicts a monopole concentration 14 orders of magnitude greater than the upper bound observed in our universe. To date, we have seen none of these exotic structures.[12]

Guth's solution to these three problems was to postulate a period of exponential expansion very early in the history of the universe. Again Greene:

> In inflationary cosmology, there was a brief instant during which gravity was repulsive and this drove space to expand faster and faster. During this part of the cosmic film, you would have to wind the film less than halfway back in order to halve the distance between the two regions. . . . the increasingly rapid separation of any two regions of space during inflationary expansion implies that halving their separation requires winding the cosmic film less – *much less* – than halfway back toward the beginning. As we go farther back in time, therefore, it becomes *easier* for any two regions of space to influence each other, because, proportionally speaking, there is more time for them to communicate. Calculations show that if the inflationary expansion phase drove space to expand by at least a factor of 10^{30}, an amount that is readily achieved in specific realizations of inflationary expansion, all the regions in space that we currently see . . . were able to communicate . . . and hence efficiently come to a common temperature in the earliest moments of the universe (Greene 2004, pp. 289–90, emphasis in original).

This inflationary period began and ended in a fraction of a second, yet a typical inflationary event could lead to 70 "e-folds." An e-fold is a logarithmic measure of how large the universe grows during an inflationary event. Here, N is the number of e-folds, and a(t) represents the scale factor of the universe at the beginning and end of inflation.[13]

$$N(t) \equiv \ln\left[a\left(t_{\text{end}}\right) / a\left(t_{\text{beginning}}\right)\right]$$

E-folds are a short hand way of expressing the huge increase in size of the universe during an inflationary event (recall Greene's factor of 10^{30}).

Hence, prior to inflation all parts of the present, observable universe could be in causal communication with one another. Inflationary expansion would also smooth the curvature of the present-day universe to be flat or nearly flat, similar to the way the curvature of a basketball would appear to vanish if it suddenly grew to the size of the Earth. Further, since our present observable universe would be only a microscopic part of the original generic manifold, the density of exotic cosmic relics would be expected to be so small that we should not see them.

Inflation was a remarkable fix to a set of serious anomalies; but it also had one more feature in store. The Hawking-Penrose singularity theorems had as one of their requirements that gravity is always attractive–just as it is for ordinary matter. But the most likely physical candidate that could account for an inflationary event was a type

of energy similar to the original cosmological constant that Einstein had proposed (Einstein 1917). This bizarre type of energy would act like repulsive gravity. This led to a philosophically desired outcome. If this "repulsive gravity" was present in the early universe and could dominate attractive gravity, then the possibility arises that the Hawking-Penrose singularity theorems did not apply to the real universe. Perhaps the universe is past eternal after all.

Inflationary theorizing eventually led to a yet grander theory, according to which the gravitationally repulsive material may, in fact, be the norm in the universe rather than the exception. Cosmologist Andrei Linde explains:

> This process, which I have called eternal inflation, keeps going as a chain reaction, producing a fractal-like pattern of universes. In this scenario the universe as a whole is immortal. *Each particular part of the universe may stem from a singularity somewhere in the past, and it may end up in a singularity somewhere in the future*. There is, however, no end for the evolution of the entire universe.
>
> The situation with the very beginning is less certain. There is a chance that all parts of the universe were created simultaneously in an initial big bang singularity. The necessity of this assumption, however, is no longer obvious.
>
> Furthermore, the total number of inflationary bubbles on our "cosmic tree" grows exponentially in time. Therefore, most bubbles (including our own part of the universe) grow indefinitely far away from the trunk of this tree. Although this scenario makes the existence of the initial big bang almost irrelevant, for all practical purposes, one can consider the moment of formation of each inflationary bubble as a new "big bang." From this perspective, inflation is not a part of the big bang theory, as we thought 15 years ago. On the contrary, the big bang is a part of the inflationary model (Linde 1998, p. 103, our emphasis).

Linde's chaotic inflation was one of two competing views for the theory. The competitor, called "new" inflation, featured the idea that there is a "false vacuum," which represents (meta)stable vacuum with a high Einstein-like cosmological constant (compared to the "true vacuum" in which we live). In "new inflation," this eternally expanding false vacuum regionally decays into the slower expanding true vacuum. It expands faster than it decays, so the process never stops.

> Chaotic inflation is rather a different idea. Here, the universe starts from a "generic manifold," which is a state of maximal entropy chaos. An energy field of different regional values pervades this manifold. Where the field is large, inflation occurs. The field can undergo quantum fluctuation to high values as well; thereby giving onset to inflation. Locally, a region will have a tendency to seek out the minimum allowed value of the energy field. This leads to a process called "reheating," which creates the ordinary matter and energy that we see around us. Meanwhile, in locations of the universe where the field energy density is high, the quantum fluctuations will tend to offset the minimizing tendency and make perpetual (globally, but not regionally) the inflationary process (Fig. 1.4).

Figure 1.4 In chaotic inflation, an initial generic manifold (a global space with random characteristics; *i.e.*, inhomogeneous and anisotropic in energy density and curvature) undergoes regional inflation. An observer sitting on an inflating spot will eventually see inflation come to a stop, but conditions are always suitable – somewhere – for inflation to proceed.

Eternal inflationary models

The 1980s and '90s witnessed a proliferation of inflationary models that theoretically allowed for a projection into an eternal past (Linde 2005). The inflationary phase mentioned above was not viewed in these models as an isolated event. Theorists began to describe the exotic energy that produces inflation as a field that pervades otherwise empty space. A key assumption was that the density[14] of the energy throughout space never changes, so that it resembles Einstein's cosmological constant. It does not depend on space or time, that is, it is constant. In that case, as space expands, more energy must continually be produced in order to maintain a constant energy density (where this energy comes from is still a matter of controversy). Space "clones" itself.[15] Occasionally, parts of this rapidly expanding space decay (convert) into the type of "empty" space that we live in. This space has a much lower-energy density, so there is now a great deal of excess energy that pervades our new "bubble." This excess energy is thought to convert into the normal matter that we see around us.

> In the latest version of inflationary theory (Fig. 1.5), these decays take the form of quantum tunneling events. Every state that possesses a positive cosmological constant is "metastable." That means that, similar to a radioactive isotope, the state lasts for a while and then changes to a different (usually lower) allowed value of the cosmological constant. This lower-energy state is initially confined to a tiny portion of space, but given that the cosmological constant makes space expand, it becomes a rapidly growing bubble that is nested within the original space.

But what happens to the original space, part of which decayed to form our universe? It is still there, continuing to expand at enormous speed. Since it (usually) has a cosmological constant larger than the new bubble, its growth outpaces that of the new bubble. Since the false vacuum expands faster than it decays, inflation is eternal

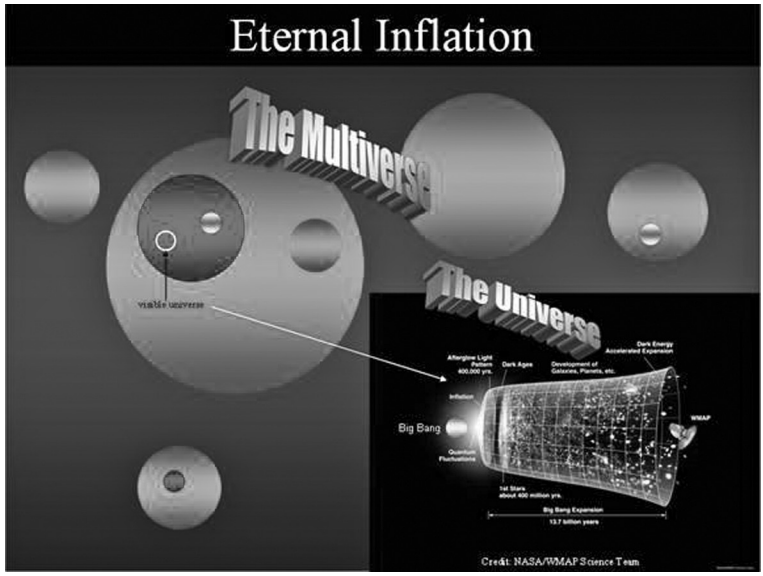

Figure 1.5 The String Landscape Inflationary Model. The Big Bang is just a regional event within a larger multiverse. There are different kinds of 'empty space' which feature different values of the cosmological constant (different colored bubbles). The larger the constant, the faster the universe expands. Our universe decayed from one of these 'false vacuum' regions.

into the future. New bubbles of low energy vacuum will continue to decay out of the expanding space.

Theorists wondered whether this process could be infinitely extended into the past. Interestingly, Guth himself, along with collaborators Alexander Vilenkin and Arvind Borde, has likely closed the door on that possibility. In 2003, Borde, Guth, and Vilenkin published an updated singularity theorem far grander in scope than the Hawking-Penrose theorems. They explain,

> Our argument shows that null and time-like geodesics are, in general, past-incomplete in inflationary models, whether or not energy conditions hold, provided only that the averaged expansion condition $H_{av} > 0$ holds along these past-directed geodesics (Borde, Guth, and Vilenkin 2003, p. 3)[16]
>
> A remarkable thing about this theorem is its sweeping generality. We made no assumptions about the material content of the universe. We did not even assume that gravity is described by Einstein's equations. So, if Einstein's gravity requires some modification, our conclusion will still hold. The only assumption that we made was that the expansion rate of the universe never gets below some nonzero value, no matter how small. This assumption should certainly be satisfied in the inflating false vacuum. The conclusion is that past-eternal inflation without a beginning is impossible (Vilenkin 2006, p. 175).

Vilenkin affirms that any universe (including universes modeled by higher dimensional cosmology, pre-Big Bang cosmology, and so forth,) which, on average, expands has to connect, in a finite time, to a past boundary (Personal communication, March 4, 2004).

Intuitively, the reason that the universe must have a beginning in the finite past is that, in an expanding space, an observer tracing out a worldline (to the future) slows down. This is the redshift. Vilenkin explains:

> Let us now introduce another observer who is moving relative to the spectators [each of whom is motionless except for the expansion of space]. We shall call him the space traveler. He is moving by inertia, with the engines of his spaceship turned off, and has been doing so for all eternity. As he passes the spectators, they register his velocity.
>
> Since the spectators are flying apart [that is, the universe is expanding], the space traveler's velocity relative to each successive spectator will be smaller than his velocity relative to the preceding one. Suppose, for example, that the space traveler has just zoomed by the Earth at the speed of 100,000 kilometers per hour and is now headed toward a distant galaxy, about a billion light years away. That galaxy is moving away from us at a speed of 20,000 kilometers per second, so when the space traveler catches up with it, the observers there will see him moving at 80,000 kilometers per second.
>
> If the velocity of the space traveler relative to the spectators gets smaller and smaller into the future, then it follows that his velocity should get larger and larger as we follow his history into the past. In the limit, his velocity should get arbitrarily close to the speed of light (Vilenkin 2006).[17]

So, looking into the past, the observer must be seen to speed up. But one cannot exceed the speed of light. The implication of this is that the past worldline of this observer has a finite length. This is the symptom of singularity; the "pathology" that Robert Wald referred to earlier. The observer will have "begun its existence a finite time ago."

The Borde-Guth-Vilenkin singularity theorem is now widely accepted within the physics community. As of this writing, it has gone largely unchallenged.[18] Instead a new round of model building has resulted based on exceptions to *this* theorem. Four alternatives present themselves (Fig. 1.6).

IIIa. Infinite Contraction

Assume that a spatially infinite universe contracted down to a singularity and then "bounced" into our present expansion. In such a case, the universe cannot be said to be on average in a state of cosmic expansion throughout its history, since the expansion phase, even if infinite, is canceled out by the contraction phase. While permissible under the Borde-Guth-Vilenkin theorem, this option is not, however, a popular option amongst contemporary cosmologists. George Ellis identifies two problems that bedevil such an approach:

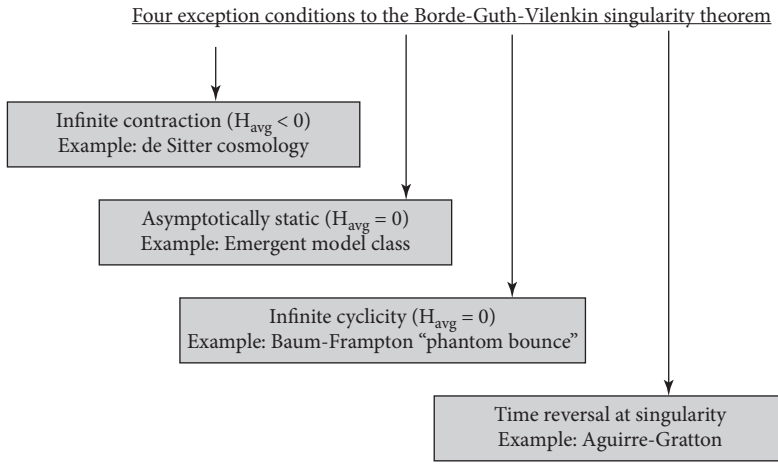
Four exception conditions to the Borde-Guth-Vilenkin singularity theorem
Infinite contraction ($H_{avg} < 0$)
Example: de Sitter cosmology
Asymptotically static ($H_{avg} = 0$)
Example: Emergent model class
Infinite cyclicity ($H_{avg} = 0$)
Example: Baum-Frampton "phantom bounce"
Time reversal at singularity
Example: Aguirre-Gratton

Figure 1.6 Post-2003 cosmological model building based on finding exceptions to the BGV theorem.

> The problems are related: first, initial conditions have to be set in an extremely special way at the start of the collapse phase in order that it is a Robertson-Walker universe collapsing; and these conditions have to be set in an acausal way (in the infinite past). It is possible, but a great deal of inexplicable fine tuning is taking place: how does the matter in widely separated causally disconnected places at the start of the universe know how to correlate its motions (and densities) so that they will come together correctly in a spatially homogeneous way in the future?
>
> Secondly, if one gets that right, the collapse phase is unstable, with perturbations increasing rapidly, so only a very fine-tuned collapse phase remains close to Robertson-Walker even if it started off so, and will be able to turn around as a whole (in general many black holes will form locally and collapse to a singularity).
>
> So, yes, it is possible, but who focused the collapse so well that it turns around nicely? (Private communication, January 25, 2006)

So there is a significant problem of *acausal* fine-tuning. One asserts not just brute contingency, but also a rather curious form of it. In the face of apparent fine-tuning, physicists usually prefer to offer some type of explanation. Consider, for example, multiverse models as an explanation of the apparent fine tuning of the fundamental physical constants, or Guth's inflationary resolution of the horizon problem (past thermodynamic equilibrium).

Second, there is the problem that the collapse becomes chaotic as it approaches the singularity. This will produce a pre-expansion start condition that is known to be dramatically different from our actual "Big Bang." This phenomenon is referred to as "BKL chaos" after its discoverers (see Belinsky, Khalatnikov, and Lifshitz 1970).[19] This

problem will appear for all attempts at a past-eternal timeline that seek to introduce a pre-Big Bang phase that "bounces" into the present expansion. In fact, the true implication of BKL may well be that it is physically impossible to "bounce" *through* a singularity.

In stating that the initial conditions "have to be set in an acausal way (in the infinite past)," Ellis puts his finger on a nettlesome philosophical issue in cosmological models featuring an infinite past, namely, they often *seem* to treat the infinite past as though it featured an infinitely distant beginning point. Several of these models are discussed in this essay. But, as we have already seen in our discussion of philosophical *kalām* arguments, such a supposition is illicit, since such an infinitely distant point is merely an ideal limit characteristic of the potential infinite, not a moment that actually once was present.[20] If we are allowed to speak of the condition of the universe at past infinity, then Zenonian paradoxes (see p. 119 in original Craig and Sinclair essay) are unavoidable.

If the past condition of the universe is acausal, then of course there was no "setting" of the condition; it just "is." Ellis is referring merely to the construction of the mathematical model. But suppose we do imagine that the boundary conditions were literally set at past infinity. Something like this was a feature of Charles Misner's old "Mixmaster" universe:

> In reality we don't expect universes to expand at exactly the same rate in every direction, and when they become asymmetrical like this they behave in a very complicated way. Although they expand in volume, one direction tends to contract while the other two expand, tending to create an expanding "pancake." But soon the contracting direction switches to expansion and one of the other two expanding directions switches into contraction. Over a long period of time, the effect is a sequence of oscillations . . . The striking thing about the sequence of oscillations of the volume of the universe as it shrinks to zero, when one runs its history back into the Big Bang at time-zero, or on into the Big Crunch at crunch-time, is that an infinite number of oscillations occur. . . .The difference between the Mixmaster Universe and Zeno's paradox is that an infinite number of physically distinct, real events happen in any finite interval of time that includes time-zero or crunch time. Measured by a clock that "ticks" on this oscillatory time, the Mixmaster Universe would be judged to be infinitely old, because an infinite number of things have happened to the past in this time, and it will "live" forever because an infinite number of things are still to happen in the future (Barrow 2005, pp. 242–43).

The Mixmaster universe is interesting in that it appears to offer a past infinite timeline that nonetheless features a clear past boundary to that timeline; that is, an infinitely distant beginning point. There *is* a question of judging the most physically appropriate measure of time. By proper time, Mixmaster arose a finite time ago from a singularity and will end its existence a finite time to the future. Time measured by oscillatory "ticks" would report a timeline that is infinite to the past and future. Barrow and Tipler elucidate:

> It is always possible to find a conformal transformation which will convert an infinite universe to a finite one and vice versa. One can always find a time coordinate in which a universe that exists for a finite proper time . . . exists for an infinite time in the new time coordinate, and a time coordinate in which a universe that exists for an infinite proper time . . . exists for only a finite time. The most appropriate *physical* time may or may not be the proper time coordinate (Barrow and Tipler 1986, p. 636).

Physicists routinely consider an infinitely distant past "beginning" point, in effect bringing infinity into their physical models through a process called a conformal transformation. Consider Barrow and Tipler, here explaining simple Friedmann-Robertson-Walker cosmological models using a device called a Penrose diagram.

> The boundaries of a Penrose diagram represent what are termed c-boundaries of the cosmological models. The c-boundaries are composed of the singularities and the points at infinity; the c-boundary of a cosmology is the edge of space-time, the "place" at which space and time begin. By convention, singularities are represented by double lines in Penrose diagrams. [For example] the initial and final singularities are the only c-boundaries in a closed Friedmann universe. An open Friedmann universe, on the other hand, has four distinct c-boundary structures: an initial singularity out of which the entire space-time arose, a single point i^0 representing spatial infinity, a 45° line $\mathfrak{I}^+$ (called "scri plus") representing "null infinity" which are the points at infinity that light rays (null curves) reach after infinite time, and a single point i^+ which all timelike curves approach for all finite times, and reach after infinite time (with the exception of those timelike curves that accelerate forever and thus approach arbitrarily close to the speed of light. These curves hit scri plus rather that i^+ at temporal infinity).
>
> A Penrose diagram allows us to define rigorously "an achieved infinity," a concept whose logical consistency philosophers have been doubtful about for thousands of years. Using the c-boundary, it is possible to discuss the topology of the "achieved infinity" and the "beginning of time" in cosmological models (Barrow and Tipler 1986, pp. 635–36).

Models like Mixmaster, the problem of supertasks[21] in general, and the meaning of conformal transformations raise the question of whether an infinite past implies the absence of a past boundary.

Fig. 1.7 shows a Penrose diagram for the universe type under consideration in this section; that is, one that contracts from infinite size down to a singularity, and then bounces into an expanding universe (see right side of diagram). Fig. 1.8 shows another type of simplified model (a de Sitter universe) that has this type of behavior. The de Sitter model includes "dark energy" while the contracting model in Fig. 1.7 includes only ordinary matter. A more realistic physical model would include both ordinary matter and "dark energy." The behavior of the universe at large size would be dominated by the dark energy, and so a de Sitter model yields good insight into

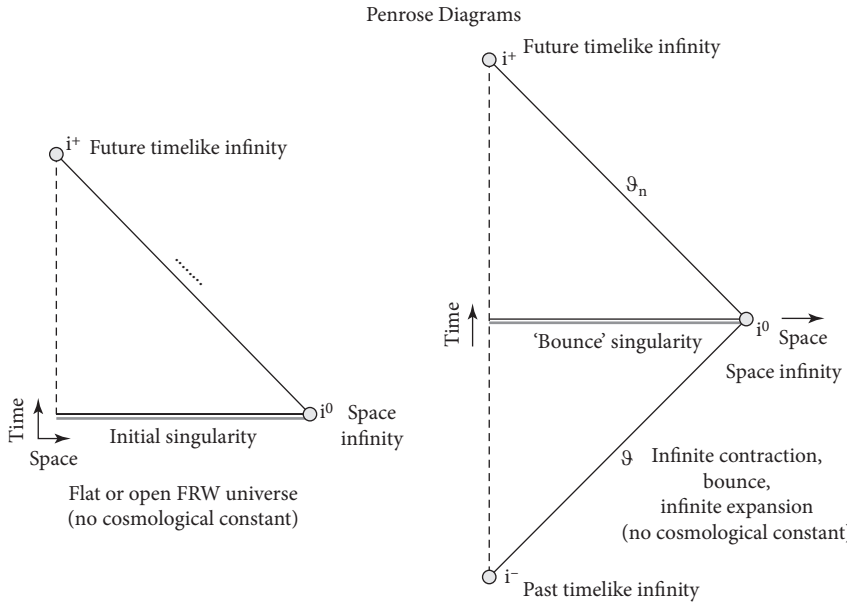

Figure 1.7 Penrose depiction of FRW cosmology.

the behavior of an infinite contraction model for the asymptotic past. By contrast, the behavior of the universe at the Big Bang would be best described by the Friedmann-Lemaître models in Fig. 1.7.

It appears there is a dilemma. One the one hand, one could have the reality of a past infinite timeline without a beginning. But then one must assert brute contingency. "Things are as they are because they were as they were."[22] Further, one must do this with respect to apparent fine-tuning. This seems implausible. One can at least say that it is unpopular, given that cosmologists are avidly seeking an *explanation* for apparent fine tuning in the form of a multiverse or a superdeterministic Theory of Everything. If we are going to give up *explanation*, then what was wrong with leaving cosmology as it was prior to 1980, namely, the standard hot Big Bang model (with associated breakdown of physics at the singularity)?[23]

The other horn involves an infinitely distant beginning point and allows the possibility of an explanation. But it opens the door for a supernatural explanation of the beginning. The *kalām* cosmological argument's second premise would be upheld, for that premise does not require that the beginning lie in the finite past.

IIIb Asymptotically static spacetime

An asymptotically static space is one in which the average expansion rate of the universe over its history is equal to zero, since the expansion rate of the universe "at" infinity is zero. Hence the universe, perhaps in the asymptotic past, is in a static

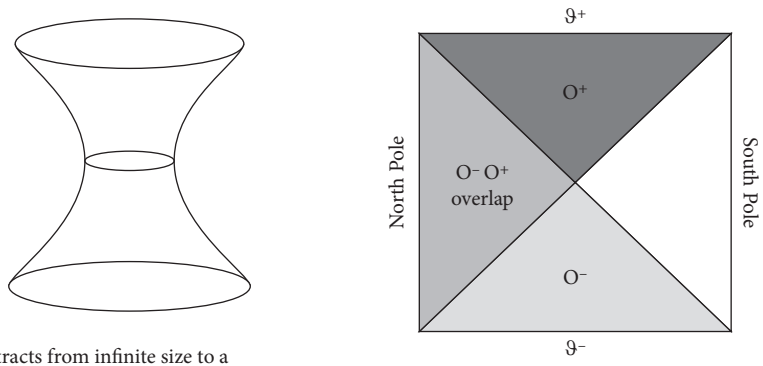
de Sitter Universe
A universe with no matter, but with a positive cosmological constant Λ
Contracts from infinite size to a minimum radius; then expands exponentially. A more realistic model would include radiation & matter which would cause a singularity at time zero, while the distant past and future would behave like the de Sitter solution.
ℑ+
O+
North Pole
O− O+
overlap
South Pole
O−
ℑ−
Penrose diagram: South & North "Pole" represent observers tracing out timelike paths. O− indicates limits of causal past of North Pole observer. O+ indicates causal future. No observer sees the full spacetime; hence the mystery of the acausal fine-tuning.

Figure 1.8 A more realistic rendering of a universe that infinitely collapses down to a Big Bang, and then expands. The de Sitter cosmology takes into account the dominant behavior of a cosmological constant for a universe of large size.

state (neither expanding nor contracting). This allows the model to escape the BGV singularity theorem.

At first blush, it would seem that the universe could hardly be said to have zero average expansion throughout its history if, as we know from observation, it has indeed been expanding! Would not the average expansion rate have to be greater than zero? No, not when we include "infinity" in the average. Consider an analogy in which the local government decides that, henceforth, everyone will pay property taxes according to the average value of property (per acre) in the county instead of on one's individual assessment. This might be good or bad for you, depending on whether you live in the high end district. But suppose that your county suddenly expanded to include the Sahara Desert. The Sahara is worthless and big, hence the average value of property, by the square mile, dives precipitously. Further, the larger the Sahara is, the closer to zero one's property taxes will be. In the limit as the Sahara grows to infinite size, one's property taxes will go to zero. In a similar way, a zero expansion condition at infinity would have the same impact on the average expansion rate. And the BGV theorem only applies to a positive *average* expansion universe. George Ellis and his colleagues have been active in this type of model building. Models of this sort belong to what is called the "Emergent" model class. They rehabilitate Einstein's static model by postulating that the universe initially existed in such a phase and

then transitioned via an inflationary phase into the universe we see around us today. Ellis and Maartens explain:

> We show here that when K = +1 [recall the curvature parameter from Friedmann's equation] there are closed inflationary models that do not bounce, but inflate from a static beginning, and then reheat in the usual way. [Recall Guth's inflation] The inflationary universe emerges from a small static state that has within it the seeds for the development of the macroscopic universe, and we call this the "Emergent Universe" scenario. (This can be seen as a modern version and extension of the Eddington universe.) *The universe has a finite initial size, with a finite amount of inflation occurring over an infinite time in the past*, and with inflation then coming to an end via reheating in the standard way (Ellis and Maartens 2004, emphasis added).[24]
>
> As such, it is a manifestly non-singular closed inflationary cosmology that *begins from a meta-stable Einstein static state* and decays into a de Sitter phase and subsequently into standard hot Big Bang evolution (Ellis, Murugan, and Tsagas 2004, emphasis added).
>
> A second, equally intriguing, possibility is that the initial Einstein static universe is created from "nothing" by some quantum tunneling process. Indeed, finiteness of the tunneling action requires that the universe created through instantonic tunneling be closed. It is not implausible, then, that through spontaneous quantum fluctuations, *a closed universe could be created in a long lived but transient Einstein static state* which then makes a transition to a finite lifetime de Sitter and subsequent marginally closed FRW phase along the lines described above (Ibid., emphasis added).[25]

Now the question that interests us is whether the past of this model is perceived as eternal. A certain amount of ambiguity attends the answer. In some accounts (such as the above) it seems pretty clear that the Emergent models do have a beginning, namely, the Einstein static state (hereafter ESS). It is also stated explicitly that the model can be constructed with ESS occurring a finite time to the past (Ellis and Maartens 2004, section V). However, in the relevant papers ESS is usually described as asymptotically approached for past infinite time: "Here . . . we consider a universe filled with a dynamical scalar field, which is past asymptotic to an Einstein static model with a radius determined by the field's kinetic energy" (Ellis and Maartens 2004, p. 1). Some philosophers who have written on the topic have a problem with the contrived nature of the past infinity in models of this type. For example, Rüdiger Vaas characterizes the Emergent models as "soft-bang/pseudo beginning" in nature. He views the asymptotic approach toward ESS as something of a mathematical artifact (Vaas 2004, p. 18).

It is worth focusing on the issue of the instability of ESS. The Einstein static universe itself was originally viewed as past eternal. But there are obvious problems with this interpretation. The reason Einstein himself originally dropped the model was its feature of unstable equilibrium. Although, in pure non-quantum GR, one can consider a static state with worldlines that trace to negative infinite time, in reality we know that

gravity is a quantum force. As Vilenkin notes, "Small fluctuations in the size of the universe are inevitable according to the quantum theory, and thus Einstein's universe cannot remain in balance for an infinite time" (Vilenkin 2006, p. 209).[26] On the other hand, the current observable universe is demonstrably *not* in a static state. A quantum (or perhaps a thermal) fluctuation is necessary to force a transition to an expanding universe. A fluctuation is in fact necessary for the two phase model to work. But this very mechanism implies that the *initial state* is not past eternal.

The best that can be done is the latest version of the Emergent model, which uses a "low-energy" solution of loop-quantum-gravity (LQG) to make the Einstein state stable against perturbations of a limited size (Fig. 1.9). In response to the question "Is the initial state metastable and therefore finite in its lifetime?" Ellis answers that the Einstein state can persist at most for a "long" but apparently *finite* time.[27]

Loop quantum gravity theorist Martin Bojowald explains that *any* perturbation, even if not of sufficient initial size to cause the system to escape the metastable potential, will *eventually* cause the system to escape it:

> Static solutions do not evolve, and so are clearly ill-suited as a model for the Universe. *But by introducing a perturbation to a static solution, one can slightly change it and thereby start a more interesting history*. Unfortunately, the classical solution [ESS] is unstable: any disturbance grows rapidly, leaving little of the initial

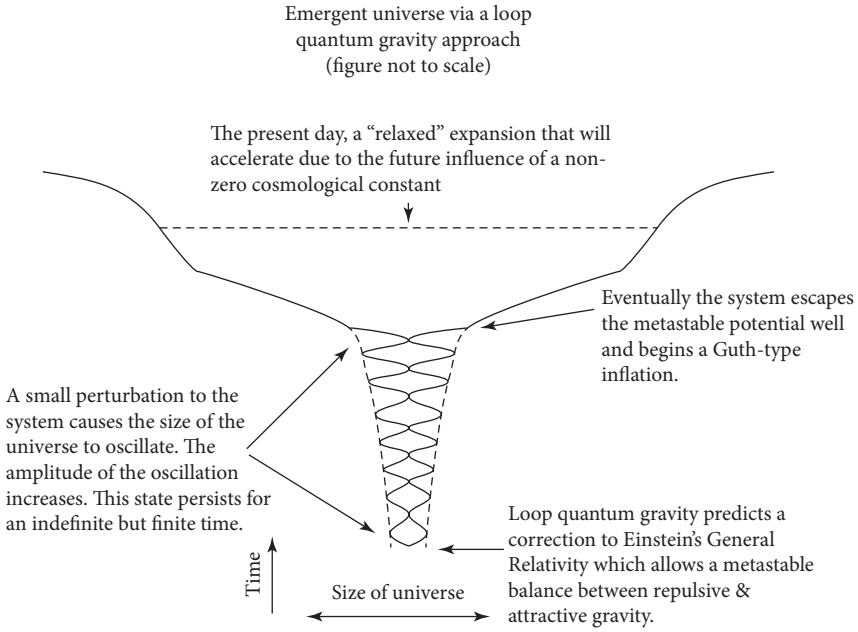
Emergent universe via a loop quantum gravity approach (figure not to scale)
The present day, a "relaxed" expansion that will accelerate due to the future influence of a non-zero cosmological constant
Eventually the system escapes the metastable potential well and begins a Guth-type inflation.
A small perturbation to the system causes the size of the universe to oscillate. The amplitude of the oscillation increases. This state persists for an indefinite but finite time.
Loop quantum gravity predicts a correction to Einstein's General Relativity which allows a metastable balance between repulsive & attractive gravity.
Time
Size of universe

Figure 1.9 Evolution of an Emergent universe from a metastable LQG state.

> state behind. The insight of Mulryne and colleagues is that quantum effects could supply all the necessary ingredients where classical solutions do not. Within the framework of loop quantum gravity, repulsion also implies static solutions at small size, but these — in contrast to the classical case — are stable. *According to the authors' model, perturbing such a state leads to small cycles of interchanging expansion and contraction*. During this process, matter will evolve slowly, and the cycles will gradually change their behavior. By itself, this perpetual recurrence and incremental change seems to lack the spark necessary for so momentous an event as the birth of the Universe. And indeed, Mulryne and colleagues identify one final theoretical ingredient that lights this spark: mediated through repulsive effects, *potential energy is gradually pushed into the matter during its slow evolution. At the point when potential energy starts to dominate kinetic energy, the mundane cycling is broken by a sudden, dramatic inflationary explosion — the emergent Universe* (Bojowald 2005, pp. 920–21, emphasis added).

Metastability implies a finite life for the associated state. Either something must have come before it or it was "created." This issue of metastability is a general problem across a wide array of model classes. Vaas elaborates:

> Meta-stable states have a local, but not a global minimum in their potential landscape and, hence, they can decay; ground states might also change due to quantum uncertainty, *i.e.*, due to local tunneling events. Some still speculative theories of quantum gravity permit the assumption of such a global, macroscopically time-less ground state (*e.g.*, quantum or string vacuum, spin networks, twistors). *Due to accidental fluctuations, which exceed a certain threshold value, universes can emerge out of that state*. Due to some also speculative physical mechanism (like cosmic inflation) they acquire – and, thus, are characterized by – directed non-equilibrium dynamics, specific initial conditions, and, hence, an arrow of time (Vaas 2004, p. 10, emphasis added).

It therefore seems that metastable (and, for that matter, unstable) states must have a merely finite lifetime. Metastable states leave unexplained how they came to exist. Universes with a metastable initial state must therefore have a beginning, consistent with the second premise of the *kalām* cosmological argument.

IIIc Cyclic universe

According to these models, the universe goes through a cycle in which it grows from zero (or near zero) size to a maximum and then contracts back to its starting condition. The universe itself is periodic, in the sense that it undergoes many such cycles, perhaps an infinite number. The average expansion of the universe would be zero in a "pure" cyclic model, since cycle by cycle one always experiences precisely equal amounts of expansion and contraction. Hence, a cyclic model evades the BGV theorem. The past is featureless. Unlike the previous two model classes, it is not the case that the universe asymptotically approaches some particular state in the infinite past.

As Vilenkin indicates, however, cyclic models face a thermodynamic problem: "A truly cyclic universe has a problem with entropy increase: it should have reached thermodynamic equilibrium by now" (Personal communication, January 19, 2007). Our observation of the present-day universe indicates that we are not at a condition of thermodynamic equilibrium–a good thing for us, as life requires non-equilibrium conditions to exist! As one looks into the past, the size of each cycle is also thought to decrease (due to radiation effect on entropy). Eventually the cycles are so small that one ends up with a different physics – which would preclude the cycling and imply a beginning to the universe.

So how does one overcome this problem? Paul Frampton and Lauris Baum have recently proposed an ingenious mechanism that breaks genuinely new ground in cosmological studies. It is surprising that they base their model on a scenario that is generally thought to imply quite the opposite of cycling. They assume that a type of dark energy pervades the universe where its equation of state (the ratio between pressure and energy density) is less than −1. This would be different from the cosmological constant mentioned earlier (equation of state equal to −1). This type of expansion is thought to lead to an event called the Big Rip. Dark energy (also called phantom energy in this context) causes the acceleration in the expansion of the universe to become so great that our visible horizon shrinks over time. Eventually this causal horizon shrinks so much that cosmological objects of smaller and smaller size become causally unbound. Galaxies, solar systems, planets, eventually even atoms get ripped apart as the expansion rate of the universe tends toward infinity. This would stop at a spatial singularity in the finite future. Baum and Frampton propose a "mosaic" model[28] to overcome the problem of entropy buildup in a single universe:

> We consider a model where, as we approach the rip, expansion stops due to a brane contribution just short of the big rip and there is a turnaround time $t = t_T$ when the scale factor is deflated to a very tiny fraction (f) of itself and only one causal patch is retained, while the other $1/f^3$ patches contract independently to separate universes. Turnaround takes place an extremely short time ($<10^{-27}$ s) before the big rip would have occurred, at a time when the Universe is fractionated into many independent causal patches (Baum and Frampton 2007, p. 1).

What happens in the Baum-Frampton approach is that very close to the Big Rip event, the universe splits into non-interacting (causally disconnected) patches. The universe has expanded so much at this point that nearly all of these patches are empty of (normal) matter and radiation. They contain only phantom[29] energy. It turns out that the entropy content of the universe (which is what interferes with cycling) is contained within the thinly spread matter and radiation. Those patches that contain only phantom energy are supposed never to undergo the Big Rip. Instead they separately undergo a *deflation* event; contracting an amount exactly equal to the expansion that the universe experienced since the Big Bang (thereby avoiding the implications of the BGV theorem). Prior to reaching a singularity, the contracting patch rebounds due to the effects of phantom energy. It then repeats the same cycle endlessly. Every patch that

undergoes this fractionates into new "universes" that themselves begin to propagate (think of dandelions spreading over your lawn). Hence the Baum-Frampton model is said to feature an infinite multiverse in addition to a beginningless cyclic behavior (Fig. 1.10). But is the model viable?

Several challenges remain to be addressed if the model is to be a viable option. First, in order to avoid the Borde-Guth-Vilenkin singularity theorem, the average contraction must equal exactly the average expansion (for every geodesic). But how is this to be done without introducing explicit fine-tuning? There is no reason that deflation of scale factor will exactly match post-Big Bang expansion. Frampton admits,

> I have no idea why it [the BGV constraint] is satisfied because it does relate expansion with dark matter to contraction without dark matter. . . . I suspect it is not fine tuning but that may ultimately depend on one's perspective: if BGV were not satisfied, infinite cyclicity is impossible (Private communication, February 5, 2007).

The problem has not yet been addressed.[30]

Second, globally, entropy should have already grown to an infinite value. How is it that the various regions remain causally disconnected? The key factor in this model is the method for jettisoning the universe's entropy. As Baum and Frampton emphasize, if matter is retained during a contraction phase:

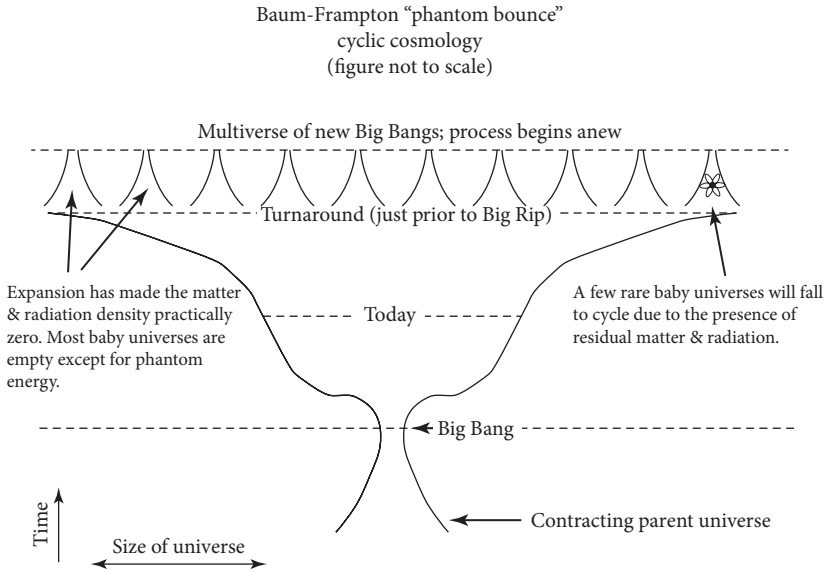
Baum-Frampton "phantom bounce"
cyclic cosmology
(figure not to scale)
Multiverse of new Big Bangs; process begins anew
Turnaround (just prior to Big Rip)
Expansion has made the matter & radiation density practically zero. Most baby universes are empty except for phantom energy.
Today
A few rare baby universes will fall to cycle due to the presence of residual matter & radiation.
Big Bang
Contracting parent universe
Time
Size of universe

Figure 1.10 Baum-Frampton phantom bounce model.

> . . . presence of dust or matter would require that our universe go in reverse through several phase transitions (recombination, QCD and electroweak to name a few) which would violate the second law of thermodynamics. We thus require that our universe comes back empty! (Baum and Frampton 2007, p. 4)

Now, globally, over infinite past time, the model posits that an infinite amount of matter and radiation (hence infinite entropy) has been produced. How is it, then, that the entropy density avoids achieving an infinite quantity? Frampton responds:

> It is true that if we retain all the separate patches entropy continues to increase. Our key idea is to retain only one causal patch for our universe whose entropy thereby drops precipitously in what we for obvious reasons call deflation (Private communication, February 5, 2007).

The sole mention of where this entropy goes occurs in the following passage: "The old problem confronting Tolman is avoided by removing entropy to an unobservable exterior region; one may say in hindsight that the problem lay in considering only one universe" (Frampton 2007a, p. 4).

Simply shoving the entropy into other "universes" raises the question whether, given infinite time, a static space, and a countable infinity of realms within the multiverse, the realms must not eventually collide. Frampton responds:

> . . . the "causal patches" remain disjoint spawning separate universes which do not interact or collide again as they start out and remain causally disconnected. This more definite answer . . . results from our better understanding of the turnaround from subsequent technical calculations (Private communication, October 10, 2007).

In our initial communication (February 2007), Frampton indicated that the problem of collision was a valid one and would be investigated. These calculations have not yet been published, hence the issue remains troubling.

Cosmologist Xin Zhang, in a recent paper (Zhang 2007a), argues that the causal disconnection mechanism at turnaround does not work precisely because the disconnected patches do come back into causal contact. Frampton has pointed out a possible error in Zhang's critique (Frampton 2007b). This appears, however, to have been rectified in a new communication (Zhang 2007b). Zhang's (new) critique is the following: Frampton uses the following form of the modified Friedmann equation[31]:

$$H^2 = \frac{8\pi G}{3}\rho\left(1 - \frac{\rho}{\rho_c}\right)$$

Here, "H" is the Hubble parameter which is defined as the time derivative of the scale factor divided by the scale factor. Recall that the scale factor is the factor by which one multiplies the size of the universe in order to represent expansion or contraction.

> Zhang clarifies: . . . ρ_c is the critical energy density set by quantum gravity, which is the maximal density of the universe. Such a modified Friedmann equation with a phantom energy component leads to a cyclic universe scenario in which the universe oscillates through a series of expansions and contractions. In the usual universe [that is one governed by the ordinary Friedmann equation], the phantom dark energy leads to a "big rip" singularity; however, in this peculiar cyclic universe, the big-rip singularity can be avoided because when ρ reaches ρ_c the universe will turn around due to [the modified Friedmann equation] (Zhang 2007b).

So the rate of change of the size of the universe is governed by the density of this special type of "phantom" energy. Zhang continues:

> When the universe approaches the turnaround point ($\rho \to \rho_c$), we have $H \to 0$. Therefore, obviously, at the turnaround, we have $H{-}1 \to \infty$. This implies that the Hubble radius becomes infinity at the turnaround point, because at that time the universe becomes static instantaneously (that is, it ceases to expand at the turnaround). Obviously, at the turnaround, the universe would not be fragmented into many disconnected causal patches (Zhang 2007b).

The inverse of the Hubble parameter ($H{-}1$, called the Hubble radius) governs the scale at which microphysics can act, that is, the scale of causal connection. While this is near zero as one nears a "Big Rip" event, it seems clear that as one approaches Baum and Frampton's turnaround, the Hubble radius grows again to an infinite value. This means that all the separate patches of the universe are not disconnected (that is, light signals can now propagate amongst them and allow them to interact). Thus, a subsequent collapse phase should include all of the matter and radiation then present, which essentially prevents a working cyclic scenario (since all the ordinary matter and radiation from the previous expansion would now be included). Frampton offers the following comment with respect to Zhang's criticism: "deflation must occur at a time before turnaround when the Hubble radius is small, before it reaches its minimum value ($x = 1/2$). Deflation remains a plausible conjecture which still requires further technical calculation to confirm" (Personal communication, Feb 7, 2008).

This seems difficult to square, as it would seem deflation = contraction, and turnaround is the time at which expansion ends and contraction begins. Baum and Frampton, in fact, issued a preprint in 2006 entitled "Deflation at Turnaround for Oscillatory Cosmology," in which they state:

> A key ingredient in our cyclic model is that at turnaround $t = t_T \pmod{\tau}$ our universe deflates dramatically with scale factor $a(t_T)$ shrinking to $\hat{a}(t_T) = fa(t_T)$ where $f < 10^{-28}$. This jettisoning of almost all, a fraction $(1 - f)$, of the accumulated entropy is permitted by the exceptional causal structure of the universe (Baum and Frampton 2006, p. 4).[32]

Frampton clarifies:

> The beginning of contraction is the turnaround. Deflation is where the causal patches separate and our entropy drops to zero.
>
> By the way the time difference between deflation and later turnaround is a trillion trillionth of a second or less!!!
>
> It [Hubble radius divergence at turnaround] does have the significance that each spawned universe is separately one causal patch at turnaround (Personal communication, Feb 7, 2008).

So deflation is not contraction. Instead it refers to causal disconnection. The Hubble radius *does* diverge at turnaround. But for some reason the causal horizons frozen in at deflation remain intact. This seems problematic. After all, causal horizons are not *real* physical barriers. They are observer dependent. For example, we on Earth have a causal horizon that stretches out some 46 billion light years. So does a space traveler in orbit around Alpha-Centauri. But the space traveler sees a different portion of the universe than we do here on Earth. There is not actually a physical barrier 46 billion light years away from each of us. This is distinct from the event horizon of a black hole, which *is* an objective physical barrier. How exactly do Baum-Frampton understand this causal disconnection? A study of Frampton's early work could perhaps clarify the situation:

> . . . the time when a system becomes gravitationally unbound corresponds approximately to the time when the growing dark energy density matches the mean density of the bound system. For a "typical" object like the Earth (or a hydrogen atom where the mean density happens to be about the density of water $\rho_{H2O} = 1\,g/cm^3$ since $10^{-24}\,g/(10^{-8}\,cm)^3 = 1\,g/cm^3$) water's density ρ_{H2O} is an unlikely but practical unit for cosmic density in the oscillatory universe.
>
> . . . the unimaginable dark energy density at turnaround of $\rho\Lambda(t_T) > 310^{27}\rho_{H2O}$. By the time the dark energy density reaches such a value, according to the Big Rip analysis [see footnote [33]], the smallest known bound systems of particles have become unbound. Additionally the constituents will be causally disconnected meaning that if *the expansion had, instead, continued to the Big Rip* the particles could no longer causally communicate (Baum and Frampton 2006, pp. 3–4, emphasis added).

This is the key. If the density of the phantom energy is $\rho\Lambda(t_T) > 10^{27}\rho_{H2O}$ then *in a Big Rip scenario*,[34] the universe would be causally unbound. But we are not in a Big Rip scenario. Instead, according to Baum and Frampton the laws of physics have changed (we now have a *modified* Friedmann equation) where *something* (perhaps extra-dimensional brane dynamics) acts to stop the expansion and leads to a contraction. In this case, it is hard to see why Frampton and Takahashi's analysis of causally unbound systems still applies. It is the runaway expansion in the Big Rip scenario that leads to and *maintains* shrinking causal horizons. If that stops, and even reverses, then it would appear reasonable to assume that the causal horizon grows along with it.

Frampton and Takahashi's explanation seems akin to a personal causal horizon as opposed to some physical barrier that results in permanent causal disconnection. What is lacking is a discussion of what causes "turnaround" and maintains causal disconnection. Without those details, it seems reasonable to entertain Zhang's misgiving concerning the viability of the universe-fractionating mechanism.

We should note Frampton's emphasis on the limited time between deflation, turnaround, and contraction. That would make the relevant issue the behavior of phantom energy right at turnaround (given that there is no time for ordinary matter and radiation to reestablish contact, given the amount of contraction that occurs in the first fraction of a second after turnaround). The question seems to be: given the *homogeneity* of the phantom energy at turnaround, and its ability to interact with its surroundings (given the unbounded Hubble horizon right at turnaround), why would the universe split into separate domains rather than precipitate a single, global contraction?

Third, the presence of *any* matter or radiation (during contraction) will prevent cycling. This could be a problem, given that spontaneous structure can form as thermal fluctuations (even if the contraction stage begins without any matter or radiation). Bousso and Freivogel explain:

> In a long-lived vacuum with positive cosmological constant, structure can form in two ways. Structure can form in the conventional way (through a period of inflation followed by reheating), or it can form spontaneously as a rare thermal fluctuation. Because de Sitter space is thermal, if the vacuum is sufficiently long-lived spontaneous structure formation will occur (Bousso and Freivogel 2007, p. 4).

The Baum-Frampton space is not a de Sitter space, but it is also thermal. Hence one would expect matter would still fluctuate into existence spontaneously. If so, then (at a reasonable probability) Baum-Frampton cyclicity would not work.

Cosmologist Thomas Banks contends that a contracting space filled with quantum fields will have an "ergodic" property as the space shrinks. Its fields become highly excited as one approaches the end of contraction and these fields will produce chaotic fluctuations. Spontaneously created matter with a different equation of state will dominate the energy density. That, and the inhomogeneity of the fluctuations, will prevent cycling. Banks and Fischler even suggest that the fields will spontaneously produce a dense "fluid" of black holes leading to a condition they call a "Black Crunch" (Banks and Fischler 2002) for arbitrary states approaching full contraction.[35] Hence it appears that Baum-Frampton cyclicity will not work.[36]

While phantom-bounce cosmologies (such as Baum-Frampton and Xin Zhang's own model) do represent a frontier worth exploring, there seem to be unanswered questions as to the viability of such an approach. The field is too young to pass full judgment. But some questions that *can* be answered (such as the ergodic/chaotic approach to a singular bounce) seem to indicate that problems native to cyclic cosmologies remain.

IIId A fourth alternative?: Time deconstruction

As Borde *et al.* point out in their seminal paper, one of their primary assumptions was the following:

> The intuitive reason why de Sitter inflation cannot be past-eternal is that, in the full de Sitter space, exponential expansion is preceded by exponential contraction. Such a contracting phase is not part of standard inflationary models, and does not appear to be consistent with the physics of inflation. If thermalized regions were able to form all the way to past infinity in the contracting spacetime, the whole universe would have been thermalized before inflationary expansion could begin. In our analysis we will exclude the possibility of such a contracting phase by considering spacetimes for which the past region obeys an averaged expansion condition, by which we mean that the average expansion rate in the past is greater than zero: $H_{av} > 0$ (Borde, Guth, and Vilenkin 2003, p. 1).

In his 2003 lecture at the Kavli institute at UCSB Guth acknowledges, "[Anthony] Aguirre and [Steve] Gratton have proposed a model that evades our theorem, in which the arrow of time reverses at the t = -infinity hypersurface, so the universe 'expands' in both halves of the full de Sitter space."[37] It is possible, then, to evade the BGV theorem through a gross deconstruction of the notion of time. Suppose one asserts that in the past contracting phase the direction of time is reversed. Time then flows in both directions *away* from the singularity. Is this reasonable? We suggest *not*, for the Aguirre-Gratton scenario (Aguirre and Gratton 2002) denies the evolutionary continuity of the universe which is topologically prior to *t* and our universe. *The other side of the de Sitter space is not our past*. For the moments of that time are not earlier than *t* or any of the moments later than *t* in our universe. There is no connection or temporal relation whatsoever of our universe to that other reality. Efforts to deconstruct time thus fundamentally reject the evolutionary paradigm.

Section III Summary

Primarily, attempts to overcome the new singularity theorem of Borde, Guth, and Vilenkin center on generating universe models that do not feature average positive expansion in their past. This can be done by having average negative expansion (that is, contraction) or by having zero average expansion (an asymptotically static model). Both attempts seem to encounter insurmountable difficulties. The contraction model features acausal fine-tuning to its asymptotic past and BKL chaos as the contraction nears a singularity in its pre-Big Bang "bounce" into our present expanding reality. BKL chaos may, in fact, prove that it is impossible for a universe to pass through a singularity.

Zero average expansion models are usually constructed in two different ways. Either the expansion asymptotically approaches zero as time (looking backwards)

approaches negative infinity. One can also consider an infinite number of cycles where expansion and contraction exactly cancel for each cycle. The first case has the dilemma that it must begin static and then transition to an expansion. Hence, the static phase is metastable, which implies that it is finite in lifetime. The universe begins to exist.

Cyclic models usually fail due to the necessary buildup of entropy from cycle to cycle. If there were an infinite number of cycles, why is the universe not in a state of "heat death"? The entropy would also have the effect of making the amplitude of the cycles (the maximum size of the universe during a cycle) grow over time. Looking backwards in time, then, this implies a first cycle to the finite past. One can attempt to overcome these problems, as Baum and Frampton do, by claiming that a single universe fractionates into a multiverse of contractions. Most of the contracting "children" from the parent universe will have shed the entropy developed from the preceding phase and hence permit cycling.

But as we have seen, their mechanism appears to fail to fractionate. Even if it did fractionate into separate entropy-free domains, even an initially empty daughter universe would develop BKL chaos when approaching a contraction and disrupt the scenario!

The last gambit, that of claiming that time reverses its arrow prior to the Big Bang, fails because the other side of the Big Bang is *not* the past of our universe.

Hence these models either have a beginning or are not viable.

IV Quantum gravity

A final, expected exception to the Hawking-Penrose theorems is a formulation known as "quantum gravity." The two great pillars of twentieth century physics, Einstein's GR (the science of the very large) and Quantum Mechanics (QM) (the science of the very small), both enjoy overwhelming observational support. However, if the standard Big Bang theory were correct in its prediction that the universe must approach a singularity in its distant past, then QM must eventually govern it. But GR is a classical theory, not a quantum field theory. For an extremely high density early universe, where gravity acted both as the dominant force and as a quantum force, we must have a new theory – quantum gravity – to describe it. Further, in 2002,[38] the first observational evidence that gravity is indeed a quantum force was reported. One of the Hawking-Penrose assumptions for their singularity theorems, however, was that GR is a correct description of the universe. Since this is not true at the scale of singularities, perhaps singularities did not exist after all. Perhaps, extrapolating backwards, the universe evolves smoothly *through* the Big Bang to an unknown past.

Here are three prominent candidates for a theory of quantum gravity (Fig. 1.11). The main job of some quantum gravity models is to get one through the singularity to an (it is hoped) eternal past. Others will accept a "beginning" to the universe but will deconstruct the notions of time, nothingness, or causation.

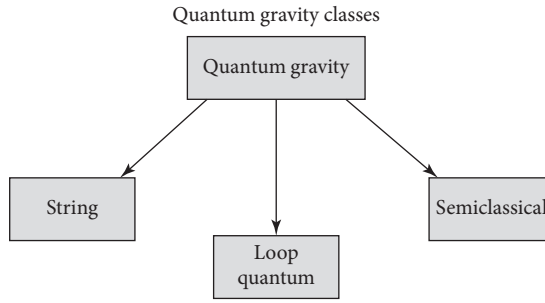

Figure 1.11 Families of quantum gravity cosmologies.

IVa. String Models

String theory is by far the most popular method proposed so far to unify quantum theory with GR. Essentially, string theory proposes that the elementary entities of nature are not point particles (zero-dimensional objects), but are strings (one-dimensional objects). String theory eliminates many of the problems that occur in particle theories. Particle interactions can occur down to literally zero distance; where force laws blow up (recall, for example, that gravity's dependence on distance is 1/r2; that is, the force becomes infinite as range goes to zero) and predict infinite (that is, nonsensical) answers. String theory calms this behavior by introducing a minimum distance (the "Planck" distance) to interactions. By "spreading out" string interactions, infinities are avoided.

Another advantage of string theory is that it can explain in a non-*ad hoc* manner the existence of different types of elementary "particles." Differing "particle" properties could be merely different types of vibrations that occur on a string (similar to musical notes). It was hoped (and still is by some) that the theory would naturally predict the characteristics of the elementary particles that are otherwise free parameters in the earlier theory (which is called the "Standard Model").

The "minimum distance" feature of string theory is thought to be a desirable feature for cosmological models because, analogously to the case of particle interactions, the standard Big Bang model predicts that the scale factor of the universe shrinks to literally zero size. String theory could "calm" this feature of the model by suggesting a minimum size. This could even overcome the Hawking-Penrose theorems and suggest that there was a "before" to the singular condition. Perhaps the Big Bang was not an ultra-mundane event at which time itself came into being. If it was not, the door is open to a past eternal universe.

IVa (i). Pre-Big Bang Inflation (PBBI)

There are two string models that can aspire to describe adequately a past infinite pre-Big Bang environment. These are the PBBI scenario of Gabriele Veneziano and Maurizio

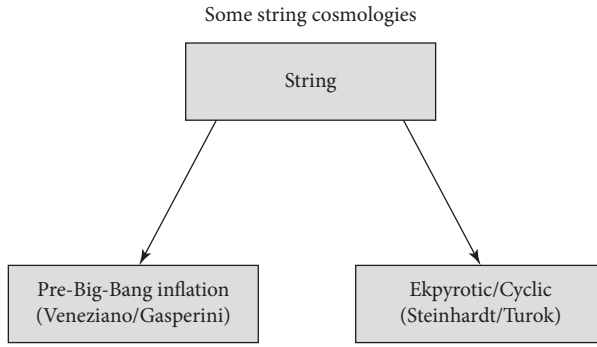

Figure 1.12 String cosmology models and proponents. String theory has also given birth to "brane" cosmology, where the emphasis is on the background within which strings propagate rather than the strings themselves. These backgrounds, called *n*-branes, can be variable in the number of their dimensions. For example, it is normally proposed that our three dimensional space is one of these '3-branes' which may or may not be floating around in a higher dimensional space (usually called "the bulk").

Gasperini and the Ekpyrotic/Cyclic model of Paul Steinhardt and Neil Turok. The pre-Big Bang inflation scenario is a classic version of an asymptotically static model. Here is how one of its authors describes it:

> According to the scenario, the pre-bang universe was almost a perfect mirror image of the post-bang one. If the universe is eternal into the future, its contents thinning to a meager gruel, it is also eternal into the past. Infinitely long ago it was nearly empty, filled only with a tenuous, widely dispersed, chaotic gas of radiation and matter. The forces of nature, controlled by the dilaton field, were so feeble that particles in this gas barely interacted. As time went on, the forces gained in strength and pulled matter together (Veneziano 2004, p. 63).

Through gravitational contraction, regions of the pre-Big Bang universe turned into black holes. Due to quantum effects, once the density of a black hole reached a critical value, it underwent a "bounce" into a Big Bang. Our universe then persists within this otherwise closed surface (with respect to the "outer" background space where everything began).

In his popularization in *Scientific American*, Veneziano seems to suggest that his beginning is an infinitely distant but never reachable (that is, ideal) point. The article implies that the model is to be interpreted realistically.[39] But one must be wary in interpreting the infinity of the past for this model, as it is easy to misuse the concept of an infinite limit, as well as to ignore the distinction between a realist and an instrumentalist interpretation of the model. The problems in interpreting this model are similar to those encountered when assessing the Emergent model class. It is worth

noting that the co-author of the PBBI model Maurizio Gasperini indicates that the entire asymptotic past (or future) should not be taken as real:

> . . . I find misleading to talk of a "future of the PBB scenario," because the PBB scenario only (possibly) applies to describe some (more or less extended) portion of the past history of our cosmos, and is expected to smoothly join the standard cosmological scenario at an early enough epoch, so as to reproduce standard results on nucleosynthesis, baryogenesis, structure formation, and so on. In other words, the PBB scenario can be regarded as a model for explaining the initial conditions of our standard cosmological configuration, in a way which is string-theory consistent, but it cannot be extrapolated towards the future without further assumptions, which at present have not been fully worked out (with the exception of the dilaton model of dark energy proposed by Piazza, Veneziano and myself on PRD 65, 023508 (2002))" (Personal communication, January 10, 2006).

Can one build a realist interpretation of this model? It is interesting to contrast the depiction of PBBI as proving "the myth of the beginning of time" in a setting where sensational conclusions are encouraged (*Scientific American*) with the characterization of PBBI as a "toy model" in a setting where scientists are naturally conservative (peer-reviewed academic journals).[40]

As described in the academic literature, the model appears to have an initial phase. The relevant phases are:

1. A static (Milne) universe, or string perturbative vacuum (SPV) phase. This means that the universe is empty (energy and energy density is zero) and is static, that is, neither expanding nor contracting globally or locally.
2. A quasi-Milne phase, which constitutes a "perturbed" string perturbative vacuum. Here "H is (small and) positive in the String frame, (small in modulus and) negative in the Einstein frame, and tends to zero as t goes to minus infinity, and the Universe approaches the SPV configuration (where H is identically zero since the spacetime is flat)" (Personal communication, Maurizio Gasperini, January 16, 2007).
3. An "inflationary" phase. In one set of coordinates (the Einstein frame), matter collapses into trapped surfaces, or black holes. In another set of coordinates (the string frame), this can be viewed as a spatial expansion. This happens regionally rather than globally.
4. A post-Big Bang FRW phase that is typical of the standard Hot Big Bang model.

The authors begin building the model at a finite time in the past where a condition called "asymptotic past triviality" (APT) obtains (Veneziano and Gasperini 2002, p. 54). APT represents the boundary between phase (2) and phase (3). The period of contraction (inflation in the string frame) is itself finite.

The authors then project this state into the future, and asymptotically evolve it into the past. Similar to the Maartens version of the Emergent model, one takes the APT past-directed duration of phase (2) to be infinite.

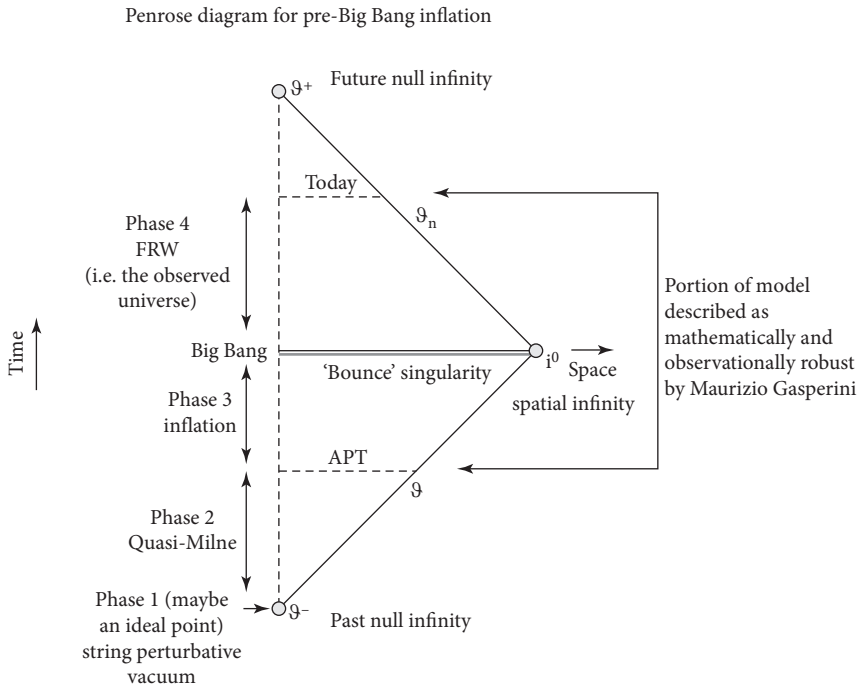

Figure 1.13 Construction of the PBBI model is given at the 'Asymptotic Past Triviality', or APT point. From here the model is projected forward in time to give the Big Bang universe. It is also projected backwards to past infinity. Note that the APT data is given at a finite time in the past in the diagram. The proposed SPV era would be at 'past null infinity', or the lower point marked 'I-'. The past infinite is 'null', rather than 'timelike,' because the gravidilaton waves that embed it are massless.

Unlike the Emergent model, perturbed SPV lasts for a while and then individual patches that meet the Hawking-Penrose condition for a closed-trapped surface begin *regional* contraction (inflation in string coordinates). Should we treat SPV like ESS? Gasperini observes that a significant feature of SPV is that it is unstable:

> . . . the SPV is unstable already at the classical level . . . it [decay of the SPV] can be described as a quantum transition, but it is a process which is also classically allowed (Personal communication, January 4, 2007).
>
> . . . the instability of the SPV is similar to the instability of a classical ball placed exactly at the top of a perfectly symmetric hill. In principle, if the system starts initially in the unique equilibrium configuration, and there are no external perturbations, the system might remain static forever. In practice, however, there are physical perturbations removing the system from equilibrium, sooner or later,

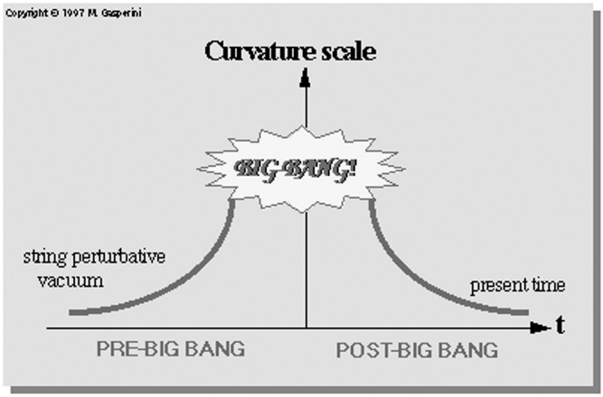

Figure 1.14 Rate of universe expansion vs. time in the PBBI Model. Source Credit: Maurizio Gasperini at http://www.ba.infn.it/~gasperin/

> with a probability which is chaotically (or randomly) distributed. In the case of the SPV the perturbations removing it from equilibrium are the quantum fluctuations of the background fields (in particular of the dilaton). In addition, the exact equilibrium configuration can only be achieved as an asymptotic extrapolation, in the limit in which the cosmic time goes to minus infinity: in practice, at any given finite physical time, the system is always displaced a bit from equilibrium (Personal communication, January 9, 2007).
>
> I would say that the SPV is not a phase extended in time, it is only an asymptotic initial state which is approached, however, in an infinite time. In practice, the physical description never starts "exactly" from that state, but from a state which represents an arbitrarily small perturbation of the SPV (Personal communication, February 27, 2007).
>
> . . . if I live in the initially collapsing portion of spacetime, then I have a chance to go through the bounce, at some epoch in the future. But this does not concern the entire spacetime. There are spacetime regions which are collapsing and eventually bouncing into a FRW like Universe, and others which do not. It is possible, in principle, to live in regions of spacetime never experiencing the bouncing nor the collapse, and staying for ever in a configuration well described by the string perturbative vacuum (or by a quantum perturbation of it) (Personal communication, January 9, 2006).

So if the SPV were real, it would be a state with a finite lifetime. Some time after decay, random portions will be sufficiently dense to form closed trapped surfaces and begin gravitational contraction. Other regions could remain indefinitely in the post-SPV state (that is, already perturbed from an equilibrium condition).

Veneziano and Gasperini's language at times suggests treating SPV as an ultimately unrealized extrapolation similar to what is described in the *Scientific American* article.

But here and elsewhere, they seem to suggest that the SPV is quite real (if the asymptotic past of the model were taken seriously):

> The whole process may be seen as the slow, but eventually explosive, decay of the string perturbative vacuum (the flat and interaction-free asymptotic initial state of the pre-big bang scenario), into a final, radiation-dominated state typical of standard cosmology (Veneziano and Gasperini 2002, p. 21).

The PBBI model is described as a "decay" or a quantum "tunneling" event similar to the semi-classical models of Hawking-Hartle and Vilenkin (see section IVc). But:

> [PBBI] can also be interpreted as a tunneling process, not "from nothing," however [which is how the Hawking-Hartle and Vilenkin models are interpreted; see section IVc], but "from the string perturbative vacuum" (Veneziano and Gasperini 2002, p. 208).

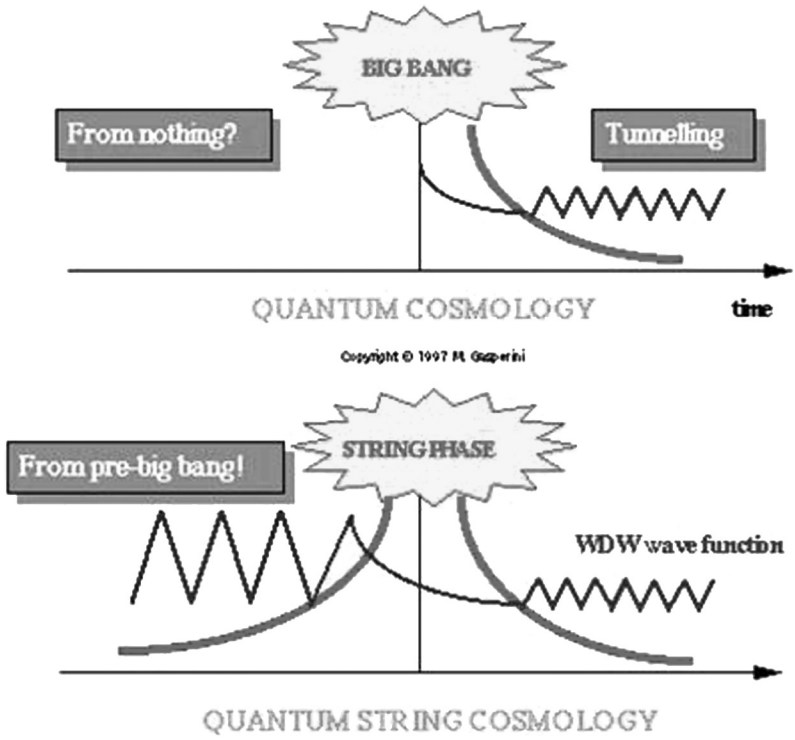

Figure 1.15 PBBI as a string quantum transition.[41] The figure contrasts "creation from nothing" models such as the Hawking-Hartle no-boundary approach or the Vilenkin "tunneling from nothing" (see section IVc) with a string approach. Here, a prior state quantum tunnels into our current FRW universe.

It is reasonable to presume that if there is a physical transition from state "A" to state 'B', and state 'B' is ontologically real, then state 'A' must be taken to be ontologically real.[42] Significant as well is that the state right after SPV decay (if SPV were a realized physical state) would obtain only a finite time ago to the past, given that Gasperini's description of the decay product is identical to the following:

> The generic regular solution thus approaches Milne as t [approaches negative infinity] but, *at any finite large* [negative time], *also contains small dilatonic (and gravitational-wave) perturbations* giving $0 < \Omega \ll 1$. As $t \to -\infty$, $\Omega \to 0$. As time goes forward, instead, Ω tends to grow until, at some critical time $-T_0$, Ω becomes O(1), in some region of space. From that moment on, in that "lucky" patch, the metric starts to deviate from Milne and dilaton-driven inflation sets in, pushing Ω extremely close to 1 in that patch (Veneziano 1998, p. 10, emphasis added).[43]

So if the SPV state were real, there could only be a finite timeline from it to the present. Only if the perturbation itself was unreal would the past eternal nature of the model survive. Is this realistic? As we look backwards from the APT point, what would the significance be of a real quantum perturbation to a state which was arbitrarily close to SPV? Would not one have a situation that was virtually identical to simply starting with "real" SPV? This is essentially a "pseudo beginning" scenario.

Vaas classifies this cosmology as "soft bang/pseudo beginning" just as he does the Emergent models.[44] Although the behavior of perturbed SPV is ultimately different from ESS (or perturbed ESS), the similarity with respect to classical (and quantum) instability should be the controlling one. The two models are mathematically similar with respect to the asymptotic approach to an infinitely distant unstable state. Vaas argues that there is no obvious arrow of time and hence the asymptotic past is likely a mathematical artifact.

Perhaps the past infinite timeline is only *technical* in nature: the duration from the APT point extrapolated *backwards* is infinite, but that is to be understood *instrumentally*. In response to the question, "if the SPV is unstable, and within a finite time quantum fluctuations of the dilaton will disturb its equilibrium, doesn't that imply that the past must be finite?" Gasperini answers,

> From a physical point of view, a spacetime manifold has an infinite (past) extension if its timelike or null geodesics can be (past) extended for infinite values of their affine parameter, without ending into a singularity. This property is satisfied by the past extension of the pre-big bang solutions, so, in this sense, they are past eternal (Personal communication, January 17, 2007).

As Gasperini indicates, it is true that, *in a sense*, the solutions are past eternal. In pure GR, it is true that backwards directed geodesics will trace all the way to past infinity. In fact, in pure GR, even if one assumed a *real* SPV state, it would still be the case that, technically speaking, backward-directed geodesics are past eternal for the same reason that a real Einstein Static State is technically past eternal.

But the relevant question seems to be this: do all past-directed geodesics from a quasi-Milne state (or a quasi-ESS) intersect with a significant system perturbation, or would at least one geodesic trace, undisturbed, to past infinity? Consider the following comparison:

> A) Suppose that the String Perturbative Vacuum (or Einstein Static State) were viewed as ontologically real rather than just idealized points at past infinity. Within a finite time, a quantum perturbation would disturb the state and the resultant timeline to the present would also be finite. The model would not be past infinite.
> B) Suppose, instead, that ESS or SPV are taken to be asymptotic ideal points. For analysis' sake, start at the present and look backwards. Within a finite time, one is arbitrarily close to the ideal condition. Now consider any (new) quantum fluctuation that occurs to the quasi-SPV (or quasi-ESS) state while one is tracing the backward timeline. The probability of this is essentially 1.

What meaningful difference is there between the universe in case A and case B (at the "new" fluctuation point)? Looking backwards, there will be an unbounded number of all types of fluctuations of all sizes, any one of which will arrest the supposed asymptotic development of the SPV (or ESS). The proposed past infinite extrapolation seems to be a mathematical artifact (or at least explanatorily vacuous). Even taken seriously (that is, not as a "toy" model), the model does not predict that the past is infinite.

IVa(ii) Ekpyrotic/Cyclic

The Ekpyrotic model is a cyclic model, but not in the old sense of a universe that undergoes an eternal periodic sequence of expansion and collapse. The Ekpyrotic model makes use of the extra-dimensional nature of string theory to propose that cycling occurs, but in a higher dimension. Authors Paul Steinhardt and Neil Turok view the new model as preferable to the post-Big Bang inflationary model because it has fewer *ad hoc* features. For example, (present day) dark energy is an add-on to the earlier model but is a natural and necessary feature of their new cycling model. String theory allows entities called "branes" that could be representative of what we would otherwise call our three (spatial)-dimensional universe. String theory demands six extra dimensions of space in order to be self-consistent. It is thought that these extra dimensions are (usually) tightly curled up around the three macroscopic dimensions and hence usually unobservable. (Think of a soda straw versus a one-dimensional line. The circle formed by looking at the straw edge-on could be thought of as a second dimension curled around the first.)

The Ekpyrotic model proposes that one of these extra dimensions (the "bulk" dimension) is of macroscopic size. Within this extra dimension lie two 3-branes, either of which could represent our universe. These 3-branes periodically collide, just as if they were connected by a spring. When they do so, the energy of the collision is transferred to the branes (Fig. 1.16).

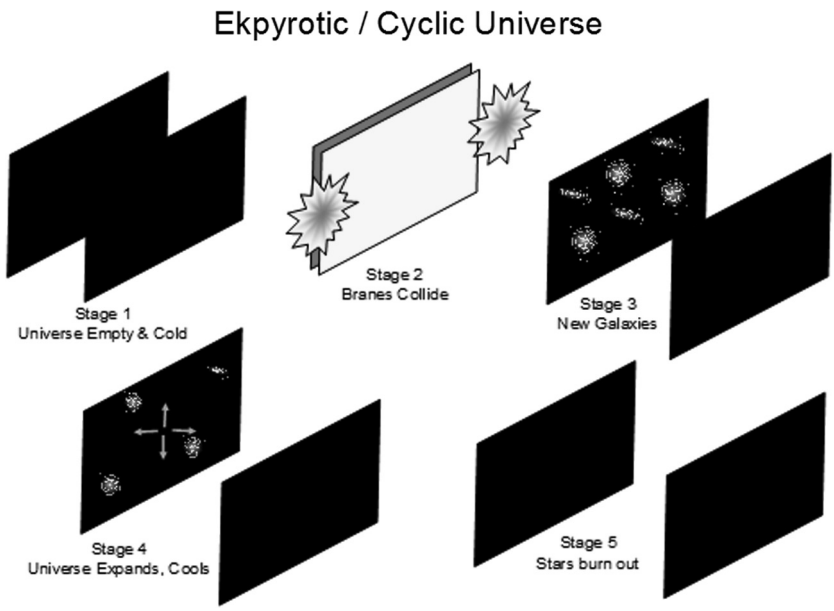

Figure 1.16 Pictorial description of Ekpyrotic cycle.

This energy is converted into the matter (and radiation) that ultimately gravitates into galaxies. The rest of the normal Big Bang sequence follows (stars, planets, and so on). As the branes separate from each other, the branes themselves are always expanding. (There are versions of the model where the branes undergo limited contraction, but always expand, on net, with each cycle.) Eventually stars burn out, the galaxies recede beyond each galaxy's visual horizon, and the universe enters a period of cold, burned out cinders. Meanwhile, the branes cycle toward another collision.

The universe is then renewed with a new cycle. The energy that is released into the branes during each collision is replenished by an inexhaustible supply—gravitational potential energy. In this way, Ekpyrosis is an "open" system. There is always a limitless supply of free energy. This feature of cosmology as an open system (as opposed to the old "heat death" scenario) is new but widespread in contemporary cosmology.

Despite avoiding the "heat death" scenario, however, there is now a new problem to solve. If new energy is dumped onto the brane for infinite time, then either the entropy density or the energy density at each point must be infinite. This would obviously be contrary to observation. The continual expansion of the infinite sized branes, however, keeps the entropy density constant. So (net) expansion is a critical model feature.

Steinhardt recognizes that his model of the universe is not truly beginningless. Here are the relevant comments from his website:

- Has the cyclic model been cycling forever?
- In principle, it is possible that the universe has undergone a semi-infinite number of cycles in its past during which the volume increases from cycle to cycle. Even though this would take an infinite time according to ordinary clocks, this cannot be the full story. This cycling regime would not cover all space-time. Something must have preceded the cycles.
- A similar issue arises in inflationary cosmology. In both cases, this is an open question. The issue is referred to as the problem of geodesic incompleteness referring to the fact that a purely expanding phase does not span the entire space-time and one has to consider what happened before.[45]

Steinhardt's website FAQ page indicates that the model is past geodesically incomplete. Here the authors comment in their published work:

> The most likely story is that cycling was preceded by some singular beginning. Consider a universe that settles into cycling beginning from some flat slice in the distant past many bounces ago. Any particles produced before cycling must travel through an exponentially large number of bounces, each of which is a caustic surface with a high density of matter and radiation at rest with respect to the flat spatial slices. Any particle attempting this trip will be scattered or annihilated and its information will be thermalized before reaching a present-day observer. Consequently, the observer is effectively insulated from what preceded the cycling phase, and there are no measurements that can be made to determine how many cycles have taken place. Even though the space is formally geodesically incomplete, it is as if, for all practical purposes, the universe has been cycling forever (Steinhardt and Turok 2005, p. 5).

Steinhardt and Turok suggest the universe began in a singularity but that "for all practical purposes" it has been cycling forever. This claim is based on the fact that virtually no information as to the initial conditions of the universe could have survived to the present. Steinhardt explains that photons carrying this information would be "semi-infinitely redshifted" (Personal communication, January 27 and 30, 2004). There is a "semi-infinite" number of cycles between that boundary and the present. How are we to understand this? The description of the model as "de Sitter-like" and the associated "semi-infinite" past timeline is nearly the same situation seen earlier with the discussion of the Misner "Mixmaster" universe (section IIIa). The key difference between Mixmaster and the Ekpyrotic model is that Mixmaster is intrinsically chaotic at its singularities (so not necessarily a good physical model of our universe), while the Ekpyrotic model avoids BKL chaos by having a positive equation of state (which is a unique feature of this cosmology) (Personal communication, January 17, 2006).[46]

Within the Ekpyrotic model an observer would see an infinite number of bounces with roughly a trillion years per bounce. So on ordinary clocks, past time is infinite (Personal communication, Jan 27 and 30, 2004). Yet there is clearly a past boundary (an infinitely distant beginning point?) preceding this behavior. One should note that

the reason this model has a beginning is precisely because it falls under the previously mentioned BGV theorem. Borde *et al.* explicitly apply their theorem to the Ekpyrotic model of Steinhardt and Turok (Borde, Guth, Vilenkin 2003).

That implies, among other things, that the "past boundary" must be reached in a finite amount of time. So Steinhardt seems to be mistaken in his prediction of a semi-infinite number of cycles, unlike the Emergent and PBBI models, which evade the BGV theorem (in proper time; that is, if "an ordinary clock" is the most appropriate measure of physical time).

Steinhardt indicated to us that measuring the distance between the present and the beginning is ambiguous. What is not ambiguous, however, is that there is a boundary. And are the infinite cycles a necessary feature of the theory? In personal communication he recognized that the theory does not require an infinite number of cycles.

Steinhardt, similar to Maurizio Gasperini, suggested to us a pragmatic view of his model. The model's description of the approach to the past boundary (and the boundary) is not (yet) rigorous. The boundary referenced by the BGV theorem could be the mark of something fundamental, such as a singularity, or instead be the marker for a transition to different physics. If one asserts that the boundary merely marks the transition to different physics, then the issue as to the age of the pre-cycling configuration asserts itself. Significantly, Steinhardt argues the cyclic solution is a dynamic attractor (this *is* rigorous) (Personal communication, January 27 and 30, 2004). What this means is that if one sets up the system with generic initial conditions within the twin-brane configuration, it will automatically, within a finite time, converge to the cyclic solution. Thus the pre-cycling configuration would have a finite lifetime. Thus there is still an unexplained origin. So the Ekpyrotic universe (whether the boundary represents a genuine singularity or not) can safely be said to begin to exist.[47]

IVb Loop Quantum Gravity

Another theory of quantum gravity is the loop quantum approach (Figure 1.17). Loop quantum gravity (LQG) takes the view that space-time itself is quantized; that is to say, space-time is divided into discrete constituent parts. It is a theory that aims to fill in the gap in standard GR by answering the question, what really happens at a singularity?

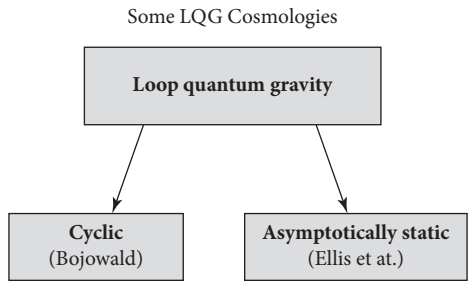

Figure 1.17 Candidate loop quantum gravity cosmologies.

According to LQG, singularities do not really exist. As in string theory, there is a minimum size to nature that prevents microscopic mathematical infinities. Hence, time and space do not come to an end as one comes to a "singularity." There will be a past timeline. This leads to the conclusion of an asymptotically static past, or a true cyclic past.

Martin Bojowald is the foremost exponent of this approach. We may think of Bojowald's model as a variation on the old Tolman cyclic model. There is only one universe. There are only three spatial dimensions. There is no "free" energy injected into the situation (such as there is in inflationary or Ekpyrotic scenarios). The Tolman model had two problems which prevented its wide acceptance: (1) There is no known physical mechanism for producing a cyclic "bounce," and (2) thermodynamic considerations show that the universe of the present day should have achieved thermodynamic equilibrium ("heat death"). This suggests that the past is finite. Bojowald recognizes both problems and believes that he can solve them. His basic approach is the same to both issues; the problems will turn out to vanish upon the generalization of current "classical" theory into LQG.

With regard to the first problem, the major difficulty has been resolving a type of chaos predicted to occur near classical singularities.[48] This chaos, named "BKL" after its discoverers (Belinsky, Khalatnikov, and Lifshitz) has been shown to be "calmed" by a loop quantum approach. As of 2007, some loop quantum theorists have been able to show that, for certain idealized models, a transition through a Big Bang condition is feasible (Ashtekar, Pawlowski, and Singh 2006). So, while a generalized proof is still lacking, the project seems promising.

The second condition is more daunting. How can there be truly cyclic behavior (one cycle looks pretty much like the last one, although there is not an event-by-event recurrence) when the second law of thermodynamics predicts that entropy must increase from cycle to cycle? Using a semi-classical approach to calculate entropy, Penrose finds that the end of our current cycle (the "Big Crunch") should differ in entropy from the Big Bang singularity by the stupendous factor of 1022 (Penrose 2005, p. 730).[49] Given no energy input from outside (and Bojowald argues that the system is truly closed), how can this outcome be avoided? There seem to be three possibilities:

(1) The problem is epistemic only. In a 2004 paper, Bojowald and his colleagues appear to favor this solution:

> While the effective dynamics is consistent with our expectations for both the beginning and the end of the universe, the apparent time reversal asymmetry remains. This is explained by the fact that the situation is, in fact, time asymmetric due to our own position in the universe. We can see only some part of it, not the whole space-time, and in particular we see only a small part of the beginning. With the current understanding, the observable part of our universe can well be part of a classical space-time with a very inhomogeneous initial singularity. Since most of the initial singularity is unobservable, however, it is not discussed further. The final singularity, on the other hand, is completely unobservable until it is reached. If we compare only observable properties from within the universe, we simply

> cannot possibly know enough to tell whether past and future singularities are similar. If we compare the theoretical structure of a space-time from outside, then we conclude that in fact there is no conceptual difference between the beginning and the end of a generic spacetime.
>
> Only if we compare the observable part of the initial singularity with the theoretical expectation for a final singularity does the time asymmetry appear (Bojowald and Hossain 2004, p. 38).[50]

2) Our current "classical" understanding of entropy is misleading. Bojowald suggested this possibility in personal communication:

> The interpretation of entropy is as a measure for the lack of information an outside observer can obtain if he just knows macroscopic parameters such as the total mass or angular momentum.
>
> The situation in non-equilibrium thermodynamics is more complicated, which would be relevant, *e.g.*, for colliding black holes or violent stages of cosmology. For cosmology, it is also not so clear what the total entropy should be associated with mathematically, so a counting is difficult. For black holes, on the other hand, entropy refers to degrees of freedom in the black hole region, which can be identified and counted within the theory.
>
> Black hole entropy then describes the lack of information in classical stages of black holes (how many quantum states there are for given mass and other parameters). This is an absolute lack of knowledge in classical gravity because the black hole region is concealed from outside observers.
>
> With quantum theory, however, black holes evaporate and thus reveal information at later stages (although it is still disputed to what degree this is realized). *The lack of information is then only temporary and apparent because an outside observer is simply not patient enough to wait until he can recover all information. In other words, entropy in this context is observer dependent and not an absolute quantity*. Since it includes only the black hole but not the observer or anything outside, it is also not the entropy relevant for cosmology.
>
> The usual intuitive picture in cosmology is as follows: When there are many black holes, this apparent entropy is very high for outside observers. *But if all degrees of freedom are considered, including those in the black holes which will re-emerge after evaporation, or one waits until after the black holes have evaporated one would obtain a smaller amount. This does not mean that entropy decreases; the accessibility of information by observers just changes* (Personal communication, February 28, 2006, our emphasis).

3) Cycle by cycle, the entropy state is genuinely reversible. Again, this alternative emerges in personal communication:

> **Sinclair**: "What it sounds to me you are saying in this last communication is that (with regard to the cyclic model) the same energy is endlessly recycled. This is a

> closed system, unlike some other cosmological proposals out there. Hence you are talking about a system that is fully reversible. This isn't a case where there are dissipative, irreversible processes that build up over time and produce the 'heat death' scenario."
> **Bojowald**: "At least in the cyclic version. If there is no recollapse at large volume, the universe would just have gone through a single bounce and will keep expanding. The end may be such a heat death, but since we don't know the field content of our universe (as evidenced by the dark matter and dark energy puzzles) the far future may be quite different from what it appears to be now" (Personal communication, February 28, 2006).

It is important to note that Bojowald (and his colleagues) are not committed to a model with a past-infinite number of cycles, as his last response shows. He is open to the possibility of an irreversible rise in entropy as a function of time. So the fact that entropy rises, cycle by cycle, and would trip up a proposed past infinite cyclic model is not, *per se,* a test of the viability of the loop quantum approach as a candidate for quantum gravity. Instead our interest (in this section) is limited to beginningless cyclic models.

Considering the first solution, Bojowald's reply to Penrose would be that there is a large, unobservable part of the initial singularity that is a genuine generic manifold, that is, a state of maximum entropy featuring random inhomogeneity and anisotropy. Hence, the entropy of the initial and final singularities would be similar. An inflation mechanism (of a small patch of this manifold) would then produce the requisite homogeneity and isotropy of the current FRW universe. Penrose, however, had anticipated this objection. Using an anthropic observer selection argument, he argues that the size of the inflationary patch we should expect to see should be much smaller based on thermodynamic criteria (by the factor 10(10^123)) (Penrose 2005, p. 763).

Penrose suggests that life might need a universe only 1/10 the size of our current (visible) universe. He obtains the probability of an appropriately sized initial patch of a generic manifold using the Hawking-Bekenstein equation for the entropy of a black hole. The exponent "123" in Penrose's formula is based on the square of the mass within the observable universe. So multiplying the radius of the universe by a tenth would have the following effect: The mass within this smaller sized universe would be reduced by a factor of 10–3 (since volume is proportional to r3), and mass is squared in the entropy formula. Hence the exponent is reduced by 6; so the overall entropy is reduced from (10^{123}) to (10^{117}). The probability of finding ourselves in either state would be approximately ten raised to the appropriate power.[51]

How many more inflationary events in a multiverse, then, would produce a life amenable but smaller universe? This is obtained from dividing out the probabilities:

$$E = 10^{-(10\hat{}17)} / 10^{-(10\hat{}123)}$$

Take the logarithm of both sides and simplify:

$$LOG(E) = -10^{117} + 10^{123} = 10^{123}$$

Here -10^{117} is negligible compared to the larger 10^{123}. Hence $E = 10^{(10\wedge 123)}$. The reciprocal of this represents the likelihood of finding ourselves *in a big universe* versus a small one. So it is exceedingly improbable to find ourselves as the product of an inflationary event of a generic manifold, as Bojowald *et al.* originally proposed. Hence, Penrose argues, the entropy of the initial manifold must be exceedingly low.

Aside from this, the epistemic argument takes no account of entropy generation during the cycle. Over infinite time, this would have to be a factor, although it may be negligible for a single cycle. Eggs break, people get old, stars burn out. This entropy may be negligible compared to black hole formation. But over infinite cycles, it would add up.

The second solution — that the "classical" understanding of entropy is misleading — mitigates the problem of entropy growth but does not resolve it. While entropy as classically calculated may be too high (that is, Penrose's estimation overestimates entropy due to contextual lack of information), the quantum approach still recognizes entropy (and entropy growth) as a genuine physical quantity. Black holes are still highly entropic. So their formation during a cycle, especially if one lands in a Big Crunch, would still cause a final manifold to have more entropy than an initial manifold. One would still expect that this situation would imply a beginning, since it implies a heat death given infinite cycles. Hence, Bojowald is only being realistic in opting for the third solution, that the cyclic LQG model does, indeed, need to be fully reversible.

But opposing the second law of thermodynamics is a formidable task; one recalls the words of early twentieth century cosmologist Sir Arthur Eddington:

> If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equations — then so much the worse for Maxwell's equations. If it is found to be contradicted by observation — well, these experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation (Eddington 1948, p. 74).

So the question is, can LQG really prove reversibility? We must await further developments in the field. It is fair to say that the prevailing view in cosmological community at large disagrees with Bojowald. As Bojowald himself admits, the jury is still out:

> **Sinclair**: "Is the assumption of entropy reversal: a) an initial assumption around which a self-consistent LQG model is built, or b) a natural fallout of LQG models?"
> **Bojowald**: "It is definitely b), as far as we can see it currently. Many details have to be filled in, but we do not make any a priori assumptions about entropy. Since entropy is not a fundamental object but a measure for our ignorance of what is happening microscopically, it is not even possible to make such an assumption in a theory like loop quantum gravity. We can only make assumptions on microscopic

> objects, and then see what this implies for more common quantities. What we don't know yet is how entropy changes balance out exactly. *So we are not sure if entropy does not increase from cycle to cycle. We can only say that the usual black hole arguments are not as strong as usually assumed*" (Personal communication, March 29, 2006, our emphasis).

To his credit, Bojowald has not simply assumed zero net entropy in his model. He is using the right approach; allowing the physics to predict entropy accrual over time. So far, however, the only reliable conclusion is that loop quantum gravity may show that Penrose's entropy arguments must be modified from their semi-classical orientation (what we have called option 2). This has failed to show, however, that universe entropy does not increase cumulatively cycle-by-cycle. The force of Penrose's argument remains intact even if his quantitative assessment of entropy must change.

Aside from the entropy issue, there remains the issue of dark energy, which may have the potential to stop cycling and induce an open-ended expansion. The current empirically observed dark energy effect, for example, appears adequate to produce an open-ended accelerated expansion. This result would be definitive if the dark energy were of the form of a cosmological constant (that is, its value were independent of space and time; see Barrow and Dabrowski 1995).[52] As related earlier (Overbye 2006), this does appear to be the fate of the present-day universe. But if an entropy gain (cycle-to-cycle) is denied, one can never have more than one "cycle." The cosmological constant would have led to open ended expansion *the first time*. Hence the initial singularity (our Big Bang) represents an absolute beginning.

If the dark energy were of the form of "quintessence" (that is, had a value that is dependent on space and/or time), however, then it would be possible that its value could reverse and be consistent with a collapse phase, even given the current observational evidence. But then a new problem could intrude. Bojowald recognizes that after the bounce and the following energy transfer, different modes of the matter fields will become excited such that the next bounce will differ from the preceding one. But if the quintessence term changes, then perhaps the most generic model of LQG would be a hybrid between the cyclic and single-bounce models. On some particular cycle a value for the quintessence term would be such that it would lead to an open-ended expansion. Bojowald responds,

> If there is just a cosmological constant, it would be fixed for all cycles and not change. But if there is some kind of quintessence, you are right that its initial conditions for the classical phase would be affected by the bounce transition. So your scenario can be realized in suitable quintessence models. However, what people usually prefer are quintessence models which have an attractor behavior at late times, or the so-called tracking solutions. This allows one to avoid too much fine-tuning, and it makes the dynamics less sensitive to changes in initial values. For such models one thus does not generically expect big changes between cycles. On the other hand, since the effect would be quite dramatic if open-ended expansion can be realized, even a non-generic possibility can be important (Personal communication, March 9, 2006).

Given an infinite number of rolls of the dice, any non-zero probability that quintessence could produce an open-ended expansion would be sufficient to do so. An open-ended expansion implies that that the overall number of cycles has been finite, and, hence the model would not be beginningless.[53]

In general, loop quantum gravity looks like a promising alternative to string theory as a candidate for quantum gravity. But building a genuinely beginningless cyclic LQG model seems to be a far more difficult challenge.[54]

IVc Semi-classical creation *ex nihilo* models

When inflation was first explored as a concept, it was understood to be a phase in the history of our universe that occurred sometime after the origin event and ended long ago. This naturally led to the question: "How could inflation itself have started?" (Figure 1.18).

That understanding has changed with the work of cosmologists such as Andrei Linde (chaotic inflation) and the recent work on the String Landscape by pioneers such as Susskind, Bousso, and Polchinski (and many others). Here, inflation is not viewed just as a phase in development, but instead as the dominant feature of a larger multiverse, within which our "universe" is just a regional phenomenon. Nevertheless, given the singularity theorems developed by Borde, Guth, and Vilenkin, inflation itself is viewed as not past eternal[55]. Hence, the old question still persists: How did inflation get started? Vilenkin explains an approach he initiated (Vilenkin 1982) to address this question:

> Many people suspected that in order to understand what actually happened in the beginning, we should treat the universe quantum-mechanically and describe it by a wave function rather than by a classical spacetime. This quantum approach to cosmology was initiated by DeWitt and Misner, and after a somewhat slow start received wide recognition in the last two decades or so. *The picture that has emerged from this line of development is that a small closed universe can spontaneously nucleate out of nothing, where by "nothing" I mean a state with no classical space and time*. The cosmological wave function can be used to calculate the probability

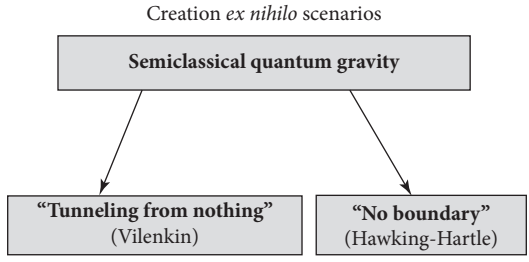

Figure 1.18 Quantum models with an explicit beginning to the finite past.

> distribution for the initial configurations of the nucleating universes. Once the universe nucleates, it is expected to go through a period of inflation, driven by the energy of a false vacuum. The vacuum energy is eventually thermalized, inflation ends, and from then on the universe follows the standard hot cosmological scenario (Vilenkin 2002, p. 2, our emphasis).

Vilenkin uses quantum tunneling of a particle through a potential well as an analogy for the whole universe. In Fig. 1.19, the portions in green represent a closed Freidmann-Robertson-Walker universe, that is, a universe which does not have enough total energy to have an open-ended expansion. Hence it goes through continued Big Bang/Big Crunch cycles. In classical physics (GR), this state of affairs would persist forever.[56] But quantum gravity provides a way out. There is a finite probability that, instead of a recollapse, the universe will "tunnel" through the energy barrier and commence an inflationary expansion instead.

This approach still does not solve the problem of creation; rather it has moved the question back one step: to the initial, tiny, closed, and metastable universe. This universe state can have existed for only a finite time. Where did it come from?

Vilenkin's solution was to consider what happens in the limit as the energy of this initial closed universe becomes zero. The scale factor of the universe becomes zero as well. This is the genesis of the claim that the universe is created from "nothing". There is no space, time, matter, or energy. This constitutes a topological transformation: "Creation of a universe from nothing . . . is a transition from the null topological sector containing no universes at all to the sector with one universe of topology S3" (Vilenkin

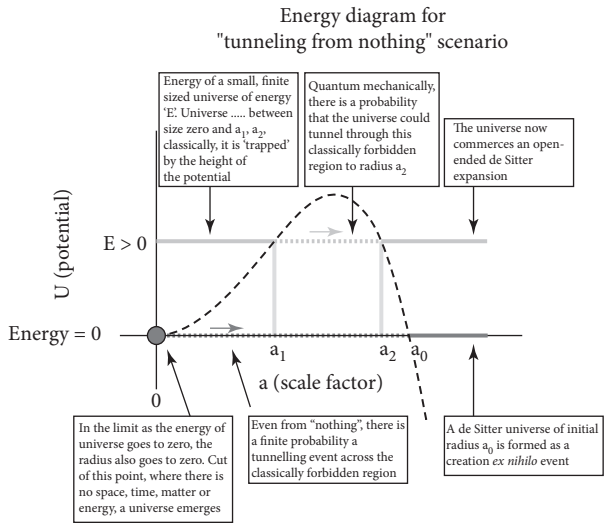
Energy diagram for
"tunneling from nothing" scenario
Energy of a small, finite sized universe of energy 'E'. Universe ..... between size zero and a_1, a_2, classically, it is 'trapped' by the height of the potential
Quantum mechanically, there is a probability that the universe could tunnel through this classically forbidden region to radius a_2
The universe now commerces an open-ended de Sitter expansion
U (potential)
E > 0
Energy = 0
a_1
a_2
a_0
a (scale factor)
0
In the limit as the energy of universe goes to zero, the radius also goes to zero. Cut of this point, where there is no space, time, matter or energy, a universe emerges
Even from "nothing", there is a finite probability a tunnelling event across the classically forbidden region
A de Sitter universe of initial radius a_0 is formed as a creation *ex nihilo* event

Figure 1.19 Creation *ex nihilo* of a universe.

1994, p. 23). Vilenkin grants that nothingness so conceived is not the same as the *absence* of being:

> I understand that a universe of zero radius is not necessarily the same thing as no universe at all. But mathematically my quantum tunneling from nothing is described by the same "shuttlecock" geometry as Hartle and Hawking [*N.B.* Fig. 1.20 below]. (In fact, the shuttlecock first appeared in my 1982 paper.) This geometry is certainly not past-eternal, and there is no point on the Euclidean sphere which you can identify as the initial universe of zero size. So, if the Hartle-Hawking approach avoids the "paradoxes of creation", I don't see why my mine doesn't (Personal communication, October 23, 2006).
>
> The mathematical description of a tunneling universe suggests that time, space and matter came into being a finite time ago. "Nothing" refers to the "prior" state (Personal communication, October 30, 2006).

The universe clearly has a beginning in this approach. But the claim seems to go beyond this: the universe "came into being" from a "prior" state of "nothing." This latter claim (that the universe comes into being) is significant, because the notion of time becomes ill-defined in these models near the beginning.

A similar approach to Vilenkin's is the "no-boundary proposal" of James Hartle and Stephen Hawking. Hartle-Hawking make use of Richard Feynman's approach to quantum mechanics. Feynman's approach (Feynman 1985) is to find the probability for a certain final quantum state by a path integral, a "sum over histories." In quantum terms, this is a superposition of states. Every possible universe history is part of the wave function; each possible end state has an associated probability of realization

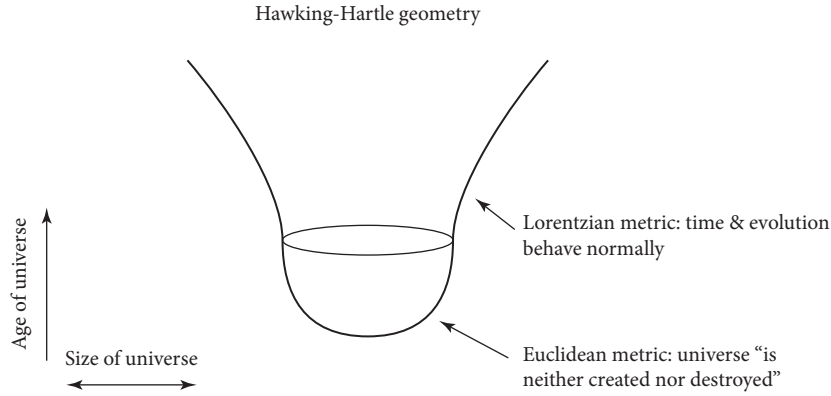
Hawking-Hartle geometry
Age of universe
Size of universe
Lorentzian metric: time & evolution behave normally
Euclidean metric: universe "is neither created nor destroyed"

Figure 1.20 Transition to "normal" time in a Hartle-Hawking approach. This is really just a mathematical technique called *analytic continuation*, which is useful for problems with "badly behaved" functions, but which does not imply ontological commitment.

(alternatively, one can view all the possible states as existing "somewhere" in a realized multiverse). The path integral of the universe, as one might imagine, presents an intractable problem; hence, Hartle and Hawking make an educated guess as to a subset of possible universes that are expected to dominate the calculation and assert compact spaces only in order to solve the problem that path (or contour) integrals often are badly behaved. They tend to diverge and give nonsense answers (like "infinity" for a probability).

To rectify the divergence problem, Hartle and Hawking make use of Euclidean metrics to describe the earliest phases of the universe's existence. This is a mathematical technique called analytic continuation (or more specifically, a Wick rotation) which allows one to analyze a better behaved function in the domain of interest. A "normal" or "Lorentzian" metric has signature (-,+,+,+), indicating a special character to the time dimension. A Euclidean metric (+,+,+,+) treats time the same as it does the space dimensions. The Wick rotation takes the real time variable "T" and replaces it with the imaginary quantity "i x T." Here "i" is the square root of negative one. Thus Hartle and Hawking are said to employ "imaginary time" in their model.

The use of a Euclidean metric has the effect of removing the initial singularity predicted by "pure" GR—hence the term "no-boundary proposal." There is, in general, no ontological commitment associated with this change in metric signature. But Hawking, in his popular writings, appears to assert the physical reality of the Euclidean signature. This is the genesis of Hawking's claim that "[the universe] would neither be created nor destroyed. It would just BE" (Hawking 1988, p. 141). Hawking goes on to ask "What place, then, for a creator?" Yet, curiously, Hartle and Hawking also claim their universe can be interpreted to have come into being out of "nothing":

> One can interpret the functional integral over all compact four-geometries bounded by a given three-geometry as giving the amplitude for that three-geometry to arise from a zero three-geometry; that is, a single point. In other words, the ground state is the probability for the Universe to appear from nothing (Hawking and Hartle, 1983, p. 2961). [They then refer to Vilenkin's "tunneling from nothing" paper.][57]

A third interpretation of these results exists as well. Recall that Gott and Li have criticized the creation *ex nihilo* approach on two grounds. (1) Transitions in quantum mechanics are always between allowed classical states (Vilenkin and Hartle-Hawking's approach has a transition from a classically forbidden region to a classically allowed region). (2) The Vilenkin and Hartle-Hawking approaches should contain realistic energy fields (something closer to what we actually see in nature). If they did, then Heisenberg's uncertainty principle would require that the initial state of their models have a finite and non-zero energy. It that is the case, then semi-classical quantum models actually start in a classically allowed metastable state, rather than "nothing." Gott and Li elaborate:

> The problem with this model [Vilenkin and Hawking-Hartle] is that it ignores the "zero-point energy." If there is a conformal scalar field ϕ, then the "energy" levels should be $E_n = n + 1/2$. Even for $n = 0$ there is a "zero-point-energy." The potential makes the system behave like a harmonic oscillator in the potential well near $a = 0$. A harmonic oscillator cannot sit at the bottom of the potential well—the uncertainty principle would not allow it. There must be some zero-point-energy and the particle must have some momentum, as it oscillates within the potential well when the field ϕ is included. Thus, when the "zero point-energy" is considered, we see that the initial state is not a point but a tiny oscillating ($0 < a < a_1$) big bang universe, that oscillates between big bangs and big crunches (though the singularities at the big bangs and big crunches might be smeared by quantum effects). This is the initial *classical* state from which the tunneling occurs. *It is metastable, so this oscillating universe could not have existed forever: after a finite half-life, it is likely to decay*. It reaches maximum radius a_1, and then tunnels to a classical de Sitter state at minimum radius a_2 where $a_2 < a_0$ (Gott and Li 1998, p. 38, our emphasis).

The relevant question for the interpretation of these models, then, is: Is the universe (i) created from nothing, (ii) neither created not destroyed but, in effect, timelessly subsistent, or (iii) left ultimately unexplained, since the existence of the initial, metastable, closed, Planck-sized universe out of which our universe was born is not itself accounted for? Option (i), if adopted, clearly implies that the universe began to exist. Nonetheless, as even Vilenkin himself recognizes, "nothing" as he describes it, is not the same as absence of being. Whatever reality underlies the laws of quantum mechanics must exist at least, so as to transform the null topological sector into an FRW universe. The laws themselves, if they exist, are mere abstract objects of a propositional nature and so do not stand in causal relations to anything. As such they are irrelevant to what happens in the world; it is the reality that corresponds to them that distinguishes merely logical from real possibility. As Heinz Pagels once remarked: "This unthinkable void converts itself into the plenum of existence—a necessary consequence of physical laws. Where are these laws written into that void? What 'tells' the void that it is pregnant with a possible universe? It would seem that even the void is subject to law, a logic that exists prior to space and time" (Pagels 1985, p. 347). Option (i), then, is mistaken, being based upon an idiosyncratic use of the word "nothing." That initial state is clearly not nothing, but something.

As for option (ii), Hawking himself seems to give good grounds for treating their proposal as an instrumental approach only. In his collaboration with Roger Penrose *The Nature of Space and Time* (Hawking and Penrose, 1996), he demonstrates the same mathematical approach (analytic continuation) to describe pair production of electron-positron pairs in a strong electric field. This is a standard mathematical technique sometimes used when complex analytic functions are better behaved in a certain domain than their real counterparts. It does not imply ontological commitment to the alternative description, however. It seems to us that given the unintelligibility of the

"imaginary time" region in these models, it is most reasonable to treat this approach as non-realist in character.

As for option (iii) we seem to have the same sort of situation that we encountered with the Emergent and PBBI models with their associated metastable ESS and SPV states. The universe cannot be past eternal because the initial metastable state can have had only a finite lifetime. This seems to us to be the most reasonable option to take for a realist interpretation of these models. It employs known, meaningful interpretations of physical phenomena from "classical" quantum theory and extends them to the quantum gravity models. One avoids the problems associated with the novelty of asserting a zero energy condition for the initial state (denied by the Heisenberg uncertainty principle), the novelty of asserting a quantum transition from a forbidden to a classically allowed state (normal quantum theory only includes transitions over or through forbidden regions from one allowed state to another), and it is consistent with more realistic energy fields. Option (iii) is also consistent with the second premise of the *kalam* cosmological argument.

2.34 Summary

Taking as a springboard challenges to the Hawking-Penrose singularity theorems, we have surveyed the historical development of three research programs each pursuing known exceptions to the theorems:[58] (1) closed time curves, (2) violation of strong energy condition (eternal inflation), and (3) falsity of GR (quantum gravity). Major theoretical developments concerning options (2) and (3) were inflationary theory (repulsive gravity), and semi-classical quantum gravity.

Closed time curves, while interesting, seem to fail given the Chronology Protection Conjecture. Counterexamples to the conjecture can be found, but they seem to be unphysical and/or infinitely fine-tuned toy models.

As for inflation, an exotic type of energy field possesses the bizarre property of *negative* pressure. This energy wants to collapse in on itself due to pressure, yet according to Einstein's equations, pressure also produces a gravitational force. If the pressure is negative, the gravitational force is *repulsive*. It turns out that the repulsion is the stronger of the two tendencies—*greatly* so. The universe can expand many orders of magnitude in a fraction of a second; from invisibly small to bigger than the entire observable sky. This development permitted a reappraisal of the question of origins that lasted two decades. Finally, a new singularity theorem, developed by Arvind Borde, Alan Guth, and Alexander Vilenkin showed that this model – the inflationary universe – still had a beginning in the finite past.

It is fascinating to note that the recent history of cosmology can be mapped by attempts to overcome these singularity theorems. Following the BGV theorem in 2003, attempts to build models have been based on exceptions to *that* theorem. These were: (1) average past expansion of universe is negative (contraction-bounce), (2) average past expansion is zero (asymptotically static universe), (3) average past expansion is zero (cyclic universe), and (4) exotic spacetime.

The first, somewhat akin to the de Sitter universe, featured an infinite contraction into a bounce at the singularity, followed by our current expansion. But it featured a Hobson's choice between an acausally fine-tuned contraction or an infinitely distant beginning point. In either case, the bounce was predicted to be chaotic (due to BKL oscillations), and hence its "Big Bang" would look nothing like the one we actually measure.

The second case, exemplified by the Emergent model class, features an unstable or metastable initial state followed by an inflationary expansion. But an unstable state (ESS) or a metastable state (the loop quantum gravity addition to the Emergent model) has a finite lifetime. So the beginning of the universe reasserts itself.

The third case has long been known, since the original models of Richard Tolman, to be problematic on entropy grounds. Baum and Frampton have sought to solve this problem through an approach dubbed "the phantom bounce." Here, the universe undergoes a superexpansion aided by the effects of "phantom" energy, which accelerates the expansion with effects similar to inflation. The universe would then fractionate into a multiverse, with almost all of the daughter universes having jettisoned their entropy given the initial expansion. Could this work? It does not appear so, the chief difficulty being that causal reconnection of universe likely occurs at the "turnaround" point (when expansion goes over to contraction) leading to *one* contraction, rather than to many. A more certain conclusion is that, even given an "empty" universe[59] undergoing contraction, chaotic fluctuations as the contraction nears a singularity would *create* matter thereby leading to a chaotic crunch, which would prevent cycling.

The fourth case features a deconstruction of time itself. It postulates two mirror-image, inflationary expansions, where the arrows of time lead *away* from a past boundary. Thus, the mirror universe is not our past. This is just a case of a double Big Bang. Hence, the universe *still* has an origin.

What about the attempts at quantum gravity models (many of which overlap with the above)? They include: (1) string models, (3) LQG models, and (3) semiclassical quantum gravity models.[60]

The most popular new field is the class of string models. Two prominent approaches are the Ekpyrotic/Cyclic and Pre-Big Bang Inflation models. But the first is subject to the BGV singularity theorem and hence has a beginning. The second is probably to be interpreted instrumentally only and seems in any case to have the characteristics of the Emergent class with respect to the initial state. Hence, it has a metastable beginning phase, called the String Perturbative Vacuum, which cannot be eternal in the past. The most popular string model is the generalization of inflationary theory known as the String Landscape. This scenario, however, is known to have a beginning to the finite past due to the same BGV theorem.

Loop quantum gravity is a competitor to string theory. We saw one such application when LQG was incorporated into the Emergent model class. Another LQG approach is to try to build a viable cyclic model. LQG seems a promising approach to address the issue of BKL chaos, hence perhaps providing a justification

for a bouncing model. But the cyclic LQG still seems to fail to account for the entropy effects that usually doom infinite cyclicity. Even if it did, current observations show that our universe is in an open-ended expansion rather than a Big Bang / Big Crunch cycle. So current LQG attempts do not appear to support a past eternal universe.

The Semi-Classical Quantum Gravity models have, in their very approach, a beginning to the finite past. The beginning has been described in three possible (and not consistent) ways. Either:

1. The universe came into being from a prior state of null topology (but somehow containing the laws of physics themselves) to a Lorentzian metric (the normal universe). Hence, the universe "tunneled from nothing" into existence.
2. The initial state of the universe is uncreated. This is due to the nature of time in a Euclidean metric. It is equivalent to a spatial dimension.
3. Because the initial state of the geometry must have a zero point energy, it is in a classical state with a Lorentzian metric. It is a metastable closed universe. Hence this state could not have existed forever and, in a manner unexplained by the model, began to exist.

The second description seems to be purely instrumental in character. The first and third imply that the universe began to exist. Hence, semi-classical models are supportive of the universe's having a beginning.

Our survey shows that contemporary cosmology is quite supportive of the second premise of the *kalām* cosmological argument. Further, this conclusion is not reached through ferreting out elaborate and unique failure conditions for scores of individual models. Rather, the repeated application of simple principles seems effective in ruling out a beginningless model.[61]

They are:

Model average expansion history	**Condition requiring a beginning**
1) Expanding models	Singularity theorems
2) Asymptotically static models	Metastability
3) Cyclic models	Second law of thermodynamics
4) Contracting models	Acausal fine-tuning

It seems that the field of cosmology therefore yields good evidence that the universe began to exist.

Notes

1 Besides the paper by Ellis *et al.*, see (Vaas 2004).
2 Hence, the temperature of the universe would be seen to rise as one looked back in time.

3 Slipher's early papers are now available on-line at <http://www.roe.ac.uk/~jap/slipher/>.

4 A geodesic is the path that a freely falling particle traces out through space and time. A timelike geodesic is traveled by a massive particle. A null geodesic is traveled by a massless particle such as the photons that make up visible light.

5 As Gott, Gunn, Schramm, and Tinsley write, "The universe began from a state of infinite density about one Hubble time ago. Space and time were created in that event and so was all the matter in the universe. It is not meaningful to ask what happened before the big bang; it is somewhat like asking what is north of the North Pole. Similarly, it is not sensible to ask where the big bang took place. The point-universe was not an object isolated in space; it was the entire universe, and so the only answer can be that the big bang happened everywhere" (Gott, Gunn, Schramm, Tinsley, March 1976, p. 65).

The Hubble time is the time since the singularity if the rate of expansion has been constant. The singularity is a point only in the sense that the distance between any two points in the singularity is zero. Anyone who thinks that there must be a place in the universe where the Big Bang occurred still has not grasped that it is space itself which is expanding; it is the two-dimensional *surface* of an inflating balloon which is analogous to three-dimensional space. The spherical surface has no center and so no location where the expansion begins. The analogy of the North Pole with the beginning of time should not be pressed, since the North Pole is not an edge to the surface of the globe; the beginning of time is more like the apex of a cone. But the idea is that just as one cannot go further north than the North Pole, so one cannot go earlier than the initial singularity.

6 However, further experiments *did* definitively establish the correspondence between GR and nature. The 1993 Nobel Prize for Physics was awarded to two astronomers: Russell A. Hulse and Joseph H. Taylor, Jr. The award was given for their study of a distant solar system consisting of a binary pulsar – two neutron stars orbiting each other. GR predicted that the orbit would shrink over time due to the emission of gravitational waves. They proved that GR is accurate to a startling degree of one part in 1014. This makes GR perhaps the best proved theory in all of physics.

7 Got indicates: "I think no one has been able to rule out CTC's. There have been no significant changes in quantum gravity since our paper. To understand whether one can create a time machine one may have to understand quantum gravity and we do not yet. Several loopholes in chronology protection have been found. Li Xin-Li and I and Cassidy, Hawking's student have found examples of quantum vacuum states that do not blow up on the cauchy horizon. Li Xin-Li's paper on the correct renormalization procedure [Phys. Rev. D. 084016 (1999)] showed that the vacuum field did not blow up for electromagnetic fields and other fields as well as for scalar fields. This means the action or entropy does not blow up either–solving a trouble Hawking and Cassidy thought existed. For matter obeying the weak energy condition instabilities are cured if the time loop occurs at the beginning of the universe, as we are proposing" (Personal communication, March 1, 2008).

8 Coule has some additional objections of his own. For example, he criticizes the nature of the vacuum, indicating that thermal fluctuations need to be of a precise form in order to avoid the radiation back-reaction described above by Thorne. This expectation is inconsistent with the Planck scale physics employed by Gott and Li.

9 In fairness to Gott and Li, it should be noted that Hiscock's criticisms are based on a semi-classical approach (an approximation of quantum gravity), and it is possible that a full theory of quantum gravity could vindicate their idea.

10 Vilenkin has also criticized the Gott-Li model (see Vilenkin 2006, p. 219). He indicates that the Gott-Li spacetime contains incomplete histories, so "This means that the spacetime itself is past-incomplete, and therefore does not provide a satisfactory model of a universe without a beginning."

11 The universe does *appear* different at various distances as we look at it. But this is due to the fact that we observe distant galaxies as they were in the past, given the time it takes for their light to reach us.

12 To understand what these exotic structures represent, consider the analogy of a pond freezing in wintertime. If the pond starts freezing in one place and the ice simply grows until it encompasses the whole pond, you will have a smooth surface. But if different parts of the pond start freezing separately, ultimately these growing "icebergs" must meet at a boundary. (Imagine taking big rocks and cracking holes in the ice; then letting it refreeze. The boundaries will be rough) The early universe was similar. These boundaries are called "defects" and can be zero, one, or two dimensional. Zero dimensional defects are called magnetic monopoles. One dimensional defects are called cosmic strings. Two dimensional boundaries are called domain walls.

13 Definition of e-fold: http://astro.uchicago.edu/~cunha/inflation/node4.html

14 In Linde's chaotic inflation, the energy field does feature quantum fluctuations that are critical to the onset of new inflationary patches.

15 Since only the "material" cause is missing, this process is an example of genuine *creatio ex nihilo* seen by physical theorists in the present day. Such recognition of efficient causation in the absence of material causation may serve to mute objections to theistic *creatio ex nihilo* as featured in the *kalām* cosmological argument.

16 H_{av} refers to the average value of the Hubble constant throughout history.

17 Alan Guth, in a 2003 lecture at UCSBs Kavli Institute says: "If we follow the observer backwards in an expanding universe, she speeds up. But, the calculation shows that if Haverage > 0 in the past, then she will reach the speed of light in a finite proper time." See http://online.kitp.ucsb.edu/online/strings_c03/guth/pdf/KITPGuth_2up.pdf

18 Andrei Linde has offered a critique, suggesting that BGV imply that all the individual parts of the universe have a beginning, but perhaps the WHOLE does not. This seems misconstrued, however, since BGV are *not* claiming that *each* past inextendible geodesic is related to a regional singularity. Rather, they claim that Linde's universe description contains an internal contradiction. As we look backwards along the geodesic, it *must* extend to the infinite past if the universe is to be past eternal. But it does not (for the observer co-moving with the expansion). Rather, past inextendible geodesics are the "symptom," not the "disease." As Robert Wald says (175Wald 1984, p. 216), "Unfortunately, the singularity theorems give virtually no information about the nature of the singularities of which they prove existence." So we don't know the nature of the singularity that the BGV theorem indicates; we know only that Linde's description of an infinite past is in error.

19 Also see Damour and Henneaux (2000): ". . . our findings suggest that the spatial inhomogeneity continuously increases toward a singularity, as all quasi-uniform patches of space get broken up into smaller and smaller ones by the chaotic oscillatory evolution. In other words, the spacetime structure tends to develop a kind of 'turbulence.'"

20 In response to the question, "Are c-boundaries [see Fig. 1.7 and following discussion for explanation of these terms] such as past and future timelike infinity and scri+ physically real edges to spacetime (real, as a black hole is an ontologically real entity) or are they merely mathematical conveniences? But if infinity is 'actual' and reachable, then a c-boundary must be an actual edge to spacetime, physically real in its ontology," Ellis responds curtly:
"1. no
2. maths - after all a spacetime diagram is just a representation of physical reality
3. in my view infinity is neither actual nor reachable."

21 A supertask is an infinite series of subtasks that can be completed in a finite time.

22 (Barrow and Tipler 1986, p. 408), attributed to cosmologist Thomas Gold.

23 See, for example, (Earman and Mosterin 1999) for a related argument.

24 Now we just showed in the previous section that Ellis has a philosophical problem with models that suggest an infinitely distant beginning point and even, in fact, the notion of a realized infinity in nature. Yet here we have a family of models developed by Ellis *et al.* that seem to suggest precisely that. This is explained via the following:

1) Infinity is so deeply ingrained in GR that pure pragmatism demands that one include the concept within one's work.
2) It is not cognitively dissonant to consider that one might be wrong and research accordingly. In fact, good scientific procedure includes an attempt to falsify one's own theories.
3) Ellis's collaborators may not have the same philosophical commitments.
4) An infinity that appears due to a technical interpretation of GR can disappear given a generalization of the theory (say, by considering Quantum Gravity).

A full look at Ellis's recent work indicates a bias *toward models with compact spaces* (*i.e.*, spatially finite either through closed curvature or topology), a skepticism with regard to infinite multiverses, and openness toward the idea of a "pseudo-beginning" in the finite past. In short, the pseudo-beginning idea is that there is timeless reality where time "switches on," producing our present state of affairs.

25 This is related to "creation from nothing" models; see section IVc.

26 We note, as well, that a perturbation to a near ESS state should be just as effective at disrupting the universe as a perturbation to a genuine ESS. Hence, a model which is only past asymptotic ESS does not escape the problem. In fact, given past infinite time, and the variety of exotic quantum universe transitions postulated throughout the cosmological literature, it seems inconceivable that any universe could possibly maintain a conserved structure over time periods "long" compared to the interval since the Big Bang.

27 He says, "note the later version of our model (astro-ph/0502589) based in the semi-classical approximation to loop quantum gravity where the static model is stable for a long while" (Private communication, January 24, 2006).

28 Mosaic model: undesirable features of a model universe may be regional in scope; consideration of a multiverse may remove those features.
29 Phantom energy: dark energy with a supernegative equation of state, *i.e.* $p/\rho < -1$.
30 In June of 2007, Paul Frampton did put a paper on the web preprint (Frampton 2007a) that partially addressed the issue. But it appears to show only that a generic cyclic model avoids the BGV theorem. That point was not in dispute. The original question still seems to persist.
31 Phantom bounce models operate on the assumption that the standard Friedmann equation is no longer valid for universe conditions where the phantom energy density is near a critical value.
32 Here τ indicates periodicy; it indicates which cycle the universe is presently in.
33 (Frampton and Takahashi 2003) and (Frampton and Takahashi 2004)
34 In a "Big Rip" the expansion rate of the universe becomes infinite and leads to a future singularity.
35 Banks complains,"I have a problem with ALL cyclic cosmologies. . . . The collapsing phase of these models always have a time dependent Hamiltonian for the quantum field fluctuations around the classical background. Furthermore the classical backgrounds are becoming singular. This means that the field theories will be excited to higher and higher energy states (define energy in some adiabatic fashion during the era when the cosmology is still fairly slowly varying, and use this to classify the states, even though it is not conserved). High energy states in field theory have the ergodic property—they thermalize rapidly, in the sense that the system explores all of its states. Willy Fischler and I proposed that in this situation you would again tend to maximize the entropy. called this a Black Crunch and suggested the equation of state of matter would again tend toward $p = \rho$. It seems silly to imagine that, even if this is followed by a re-expansion, that one would start that expansion with a low entropy initial state, or that one had any control over the initial state at all" (Private communication, Oct 12, 2007).
36 We note that Xin Zhang has his own competing cyclic model, in which he admits that thermal fluctuations pose a serious problem for phantom bounce cosmologies (Zhang, Zhang, and Liu 2007): "It is noteworthy that the cyclic universe discussed in this paper is an ideal case, and there are still several severe obstacles existing in the cyclic cosmology, such as the density fluctuation growth in the contraction phase, black hole formation, and entropy increase, which can obstruct the realization of a truly cyclic cosmology" (emphasis added).
37 Alan Guth, speech to the Kavli Institute for Theoretical Physics, October 2003, Web: http://online.kitp.ucsb.edu/online/strings_c03/guth/
38 Nesvizhevsky *et al* (2001), see http://www.newscientist.com/article.ns?id=dn1801 297, or http://physicsworld.com/cws/article/news/3525. The experiment was done by bouncing supercold neutrons and noticing that the height of the bounce was quantized.
39 Or is it *Scientific American* that has inserted the realist interpretation?
40 "The so-called 'pre-big bang' scenario described in this report has to be seen in the above perspective as a possible example, even just as a toy model of what cosmology can look like if we assume that the sought for standard model of gravity and cosmology is based on (some particular version of) superstring theory" (Veneziano and Gasperini 2002, p. 4).

41 "WDW" refers to the Wheeler-DeWitt equation, which is the basis for the semi-classical cosmologies to be discussed in section IVc.

42 If state "A" is unreal, then no transition at all takes place but rather an absolute "coming into being" of state "B." On this interpretation, the universe began to exist.

43 For clarification, a Milne universe is a special state where the critical density parameter is exactly equal to zero; the universe is empty. Recall that the critical density parameter determines if a Friedmann-Robertson-Walker cosmology will have a closed, flat, or open geometry. It is closed if < 1, flat if = 1, and open if > 1. If the cosmological constant is zero, then a closed universe will recollapse, and a flat or open universe will be ever-expanding.

44 Vaas elaborates: "A related problem refers to the pre-big bang model (Veneziano & Gasperini 2003). Here the string vacuum – where a local collapse in the Einstein frame (which corresponds to a dilaton-driven inflation in the string frame) before the big bang occurs – is quite simple, homogeneous, almost empty and does not have an overall arrow of time. But, mathematically, the origin of the pre-big bang – or, to be more precise, any pre-big bang, for the model does also imply a multiverse scenario – traces back to a maximally simple, static condition only in the infinite past (principle of asymptotic past triviality). But this can also be interpreted just as a local predecessor of a big bang and not a feature characterizing the infinite string vacuum as a whole" (Vaas 2004, pp. 18–19).

45 Cosmologist Paul Steinhardt's internet site; FAQ section for Ekpyrotic model http://wwwphy.princeton.edu/~steinh/cyclicFAQS/index.html#eternal

46 Recall that the equation of state is negative (for a cosmological constant pressure = minus density) for a typical inflationary model. For Ekpyrosis, the collision singularity occurs in the "bulk" dimension between the two branes; not within the brane that we ourselves would live on.

47 Rüdiger Vaas has suggested that, while the cycling in the Ekpyrotic model had an origin, the brane components themselves could be past eternal. As we have seen, however, an explanation is necessary as to how they got into their initial non-cycling state a finite time to the past. Vaas does not provide one (Vaas 2004, p. 17).

48 This is the same type of chaos that Ellis mentioned as an obstacle to infinite contraction models in section IIIa and is related to the problem mentioned by Banks for oscillating models (section IIIb)

49 Penrose considers that there are 10^{80} baryons in the observable universe. He then suggests that the maximum entropy for the universe is equivalent to a black hole with this mass. Should the fate of the universe be to ultimately collapse in a "Big Crunch," this would be the entropy contributed by these 10^{80} baryons. Penrose uses the Hawking-Bekenstein formula for the entropy of a black hole, and in natural units (where constants of nature such as the speed of light are set to unity), finds that this entropy is approximately 10^{123}. The entropy of our universe in the current day is far lower than this—by about 22 orders of magnitude.

50 Note that the terms "initial" singularity and "final" singularity refer to the states that begin and end our current cosmological cycle. Bojowald *et al.* are not referring to the beginning and end of time.

51. The entropy is related to the number of possible configurations of a system of particles. The number of configurations is approximately equal to the exponential of the entropy. Given the size of the numbers involved, there is essentially no difference between e^x and 10^x. So Penrose uses base 10 for convenience.
52 According to NASA, http://map.gsfc.nasa.gov/m_mm/mr_limits.html, their current data tends to favor the cosmological constant theory for dark energy as opposed to quintessence, although the latter is not ruled out.
53 Barrow and Dabrowski do indicate that if the dark energy were of the type to decay away into matter and radiation; *i.e*. if it is impermanent, then cycling would recommence after the decay.
54 We note that there are other LQG models that feature Vaas-type pseudo-beginnings and seem to us to be viable.
55 One remembers, of course, exceptions such as asymptotically static models and Linde's objection that the behavior of the whole may not be the same as its parts.
56 Assuming, of course, that the problems of cyclic cosmologies (entropy buildup, bounce physics, *etc*.) were solvable.
57 Note that in Vilenkin's comment (note), he indicates that the geometry of his model can be understood in terms of an initial Euclidean metric (which would represent the "forbidden" region of the energy diagram (Fig. 1.20).
58 The fourth and fifth conditions, *viz*., satisfaction of a generic energy condition and the existence of a closed, trapped surface in our past are easily met.
59 That is, except for the zero entropy phantom energy.
60 In interests of economy we have not discussed earlier attempts to postulate a regional, immanent status for the Big Bang which leave the origin of the initial space in question. It would seem that this space cannot contract, expand, or be static without violating criteria mentioned earlier (singularity theorems, chaotic bounces, metastable beginnings, *etc*.). Thus, these models proved to be untenable.
61 This does not exhaust the list of possible model formulations. Different types of mosaic models are likely to be the next thing to come down the pike. Nor did we survey every model that exists in the present (nor could we). There is also the complaint of philosopher James Brian Pitts: "What rights do unborn theories possess?"

References

Aguirre, A. and Gratton, S. (2002), "Steady State Eternal Inflation," in *Phys. Rev. D65,* 083507, Preprint: http://arxiv.org/abs/astro-ph/0111191.

Ashtekar, Abhay, Pawlowski, Tomasz, and Singh, Parampreet (2006), "Quantum Nature of the Big Bang," in *Phys. Rev. D* 74, 084003, Preprint: http://arxiv.org/abs/gr-qc/0602086v2.

Banks, T. and Fischler, W. (2002), "Black Crunch," http://arxiv.org/abs/hep-th/0212113v1

Barrow, John D. and Tipler, Frank (1986), *The Anthropic Cosmological Principle*. Oxford: Clarendon Press.

Barrow, John, and Dabrowski, Mariusz (1995), "Oscillating Universes," in *Mon. Not. R. Astron. Soc.* 275, 850–862.
Barrow, John D. (2005), *The Infinite Book*, New York: Pantheon Books.
Baum, Lauris, and Frampton, Paul H. (2006), "Deflation at Turnaround for Oscillatory Cosmology," http://arxiv.org/abs/astro-ph/0608138.
Baum, Lauris, and Frampton, Paul H. (2007), "Turnaround in Cyclic Cosmology," in *Phy. Rev. Lett.* 98, 071301, page 1, Preprint: http://arxiv.org/abs/hep-th/0610213.
Belinsky, V. A., Khalatnikov, I. M., and Lifshitz, E. M. (1970), "Oscillatory approach to a singular point in the relativistic cosmology," in *Adv. Phys.* 19, 525.
Bojowald, Martin, Date, Ghanashyam, and Hossain, Golam Mortuza (2004), "The Bianchi IX model in Loop Quantum Cosmology," in *Class. Quant. Grav.* 21, 3541, Preprint: http://arxiv.org/abs/gr-qc/0404039.
Bojowald, Martin (2005), "Original Questions," *NATURE Vol* 436 18 August 2005, pg. 920–921.
Borde, A., Guth, A., Vilenkin, A. (2003), "Inflationary spacetimes are not past-complete," in *Physical Review Letters*, vol. 90, p. 151301, Preprint archive: http://arxiv.org/abs/gr-qc/0110012.
Bousso, Raphael, and Freivogel, Ben (2007), "A paradox in the global description of the Multiverse," in *Journal of High Energy Physics, Issue 06*, Preprint: http://arxiv.org/PS_cache/hep-th/pdf/0610/0610132v2.pdf.
Coule, D.H. (2005), "Quantum Cosmological Models," in *Class.Quant.Grav.* 22 (2005) R125–2308. preprint: http://arxiv.org/abs/gr-qc/0412026v3, pg. 31.
Damour, Thibault and Henneaux, Marc (2000) "Chaos in Superstring Cosmology," in *Phys. Rev. Lett.* 85, 920-923, Preprint: http://aps.arxiv.org/abs/hep-th/0003139.
Davies, Paul (1978), "Spacetime Singularities in Cosmology," in *The Study of Time III*, ed. J. T. Fraser (Berlin: Springer Verlag), pp. 78–9.
Davies, Paul (1983). *God and the New Physics.* New York: Simon & Schuster.
Earman, J., and Mosterin, J. (1999), "A critical look at inflationary cosmology," in *Philosophy of Science,* 66.
Eddington, A. S. (1948), *The Nature of the Physical World*, New York: Macmillan, p. 74.
Einstein, Albert (1917). "Cosmological Considerations on the General Theory of Relativity," in *The Principle of relativity*, by A. Einstein, *et. al.*, with Notes by A. Sommerfeld, trans. W. Perrett and J. B. Jefferey (rep. ed.: New York: Dover Publications, 1952), pp. 177–188.
Ellis, G. F. R.; Kirchner, U., and Stoeger, W. R. (28 August 2003). "Multiverses and Physical Cosmology." http://arXiv:astro-ph/0305292 v3.
Ellis, G. F. R., Murugan, Jeff, and Tsagas, Christos G. (2004), "The Emergent Universe: An explicit construction," in *Class. Quant. Grav.* 27, 233, preprint: http://arxiv.org/abs/gr-qc/0307112.
Ellis, G. F. R., Maartens, Roy (2004) "The Emergent Universe: Inflationary cosmology with no singularity," *Class. Quant. Grav*. 21, 223, preprint: http://arxiv.org/abs/gr-qc/0211082.
Feynman, Richard (1985), *QED, The Strange Theory of Light and Matter*. Princeton University Press.
Frampton, P. H. and Takahashi, T. (2003), "The Fate of Dark Energy," in *Phys. Lett. B*557, 135, Preprint: http://arxiv.org/abs/astro-ph/0211544.

Frampton, P. H. and Takahashi, T. (2004), "Bigger Rip with No Dark Energy," in *Astropart. Phys.* 22, 307, Preprint: http://arxiv.org/abs/astro-ph/0405333.
Frampton, Paul (2007a), "Cyclic Universe and Infinite Past," preprint: http://aps.arxiv.org/PS_cache/arxiv/pdf/0705/0705.2730v2.pdf.
Frampton, Paul (2007b), "Comment on 'Can Black Holes be Torn Up by a Phantom in Cyclic Cosmology?' by X. Zhang," preprint: http://arxiv.org/PS_cache/arxiv/pdf/0709/0709.1630v1.pdf.
Frege, Gottlob (1968). *The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number*. Trans. J. L. Austin. 2nd rev. ed. Evanston, Ill.: Northwestern University Press.
Friedman, Michael (1973). "Grünbaum on the Conventionality of Geometry." In *Space, Time and Geometry*, pp. 217–33. Ed. Patrick Suppes. Synthese Library. Dordrecht: D. Reidel.
Friedmann, Alexander (1922), "Über die Krümmung des Raumes," *Zeitschrift für Physik* 10 (1922): 377–86; G. Lemaître, "Un univers homogène de masse constante et de rayon croissant, rendant compte de la vitesse radiale des nébuleuses extragalactiques," *Annales de la Société scientifique de Bruxelles* 47 (1927): 49–59.
Gott III, J. Richard, Gunn, James E., Schramm, David N., and Tinsley, Beatrice M., "Will the Universe Expand Forever?" *Scientific American* [March 1976], 65).
Gott III, J. Richard and Li-Xin, Li (1998). "Can the Universe Create Itself?" *Physical Review D* 58 (2): 023501–1.
Greene, Brian (2004), *The Fabric of the Cosmos*, Publisher: Alfred A. Knopf.
Guth, Alan H. (1981), "Inflationary universe: A possible solution to the horizon and flatness problems," in *Physical Review D, Volume* 23, *Number* 2, 15 january.
Hawking, Stephen and Penrose, Roger (1970), "The Singularities of Gravitational Collapse and Cosmology," in *Proceedings of the Royal Society of London, series A*, 314, pg. 529–548.
Hawking, Stephen, and Hartle, James (1983), "The Wave Function of the Universe," in *Physical Review D*, *Volume* 28, *Number* 12, 15 December, 2960–75.
Hawking, Stephen (1988), *A Brief History of Time*, Bantam Books, pg. 141.
Hawking, Stephen (1992), "The chronology protection conjecture," in *Phys. Rev.,* D46, 603–611.
Hawking, Stephen and Penrose, Roger (1996), *The Nature of Space and Time*, Princeton University Press.
Hawthorne, Jonathan. 2000. "Before-effect and Zeno causality." *Noûs* 34: 622–33.
Hiscock, William A. (2000), "Quantized Fields and Chronology Protection," preprint: http://arxiv.org/abs/gr-qc/0009061v2, pg. 3.
Hubble, Edwin (1929), "A Relation between Distance and Radial Velocity among Extra-galactic Nebulae," *Proceedings of the National Academy of Sciences* 15 (1929): 168–73.
Linde, Andrei (1998), "The Self-Reproducing Inflationary Universe," Pg. 103 of *The Magnificent Cosmos*, special edition of *Scientific American*, May.
Linde, Andrei (2005), "Inflation and String Cosmology," *J.Phys.Conf.Ser.* 24 151–160, Preprint: http://arxiv.org/PS_cache/hep-th/pdf/0503/0503195v1.pdf.
Naber, Gregory L. (1988), *Spacetime and Singularities: an Introduction*. Cambridge: Cambridge University Press, 1988, pp. 126–27.
Nerlich, Graham (1976). *The Shape of Space.* Cambridge: Cambridge University Press.

Nesvizhevsky *et al* (2001), "Quantum states of neutrons in the Earth's gravitational field" *Nature* 415, pg. 297.
Overbye, Dennis (2006), "9 Billion-Year-Old 'Dark Energy' Reported," in *The New York Times* (November 17). http://www.nytimes.com/2006/11/17/science/space/17dark.html.
Pagels, Heinz (1985), *Perfect Symmetry: The Search for the Beginning of Time*, New York: Simon & Schuster, pg. 347.
Penrose, Roger (2005), *The Road to Reality*, New York: Alfred A. Knopf.
Poincaré, Henri (1982). *Science and Hypothesis*. In *The Foundations of Science*, by Henri Poincaré. Repr. ed. Washington, DC: University Press of America.
Quinn, Phillip L. (1976). "Intrinsic Metrics on Continuous Spatial Manifolds." *Philosophy of Science* 43: 396–414.
Reichenbach, Hans (1958). *The Philosophy of Space and Time.* Trans. Maria Reichenbach and John Freund. New York: Dover.
Russell, Bertrand (1937). *The Principles of Mathematics*. 2nd ed. London: George Allen & Unwin.
Senor, Thomas D. (1993). "Divine Temporality and Creation *ex nihilo*." *Faith and Philosophy* 10: 86–92.
Smith, Quentin (2007). "Kalam Cosmological Arguments for Atheism." In *The Cambridge Companion to Atheism*, pp. 182–98. Ed. M. Martin. Cambridge Companions to Philosophy. Cambridge University Press.
Steinhardt, Paul, and Turok, Neil (2005). "The Cyclic Model Simplified," in *New Astron. Rev.* 49, pg. 43–57, preprint: http://arxiv.org/abs/astro-ph/0404480
Thorne, Kip (1994) *Black Holes and Time Warps*, W.W. Norton & Co., pg. 505–506
Vaas, Rüdiger (2004). "Time before Time: Classifications of universes in contemporary cosmology, and how to avoid the antinomy of the beginning and eternity of the world." In *Knowledge and Belief. Papers of the 26th International Wittgenstein Symposium*. Löffler, W., Weingartner, P. (eds.): Austrian Ludwig Wittgenstein Society: Kirchberg am Wechsel, pp. 351–353 preprint: <http://arXiv.org/abs/physics/0408111>.
Veneziano, Gabriele (1998), "A Simple/Short Introduction To Pre-Big-Bang Physics/ Cosmology," preprint: http://arxiv.org/abs/hep-th/9802057.
Veneziano, Gabriele (2004), "The Myth of the Beginning of Time," May *Scientific American*, pg. 63.
Veneziano Gabriele, and Gasperini, Maurizio (2002), "The Pre Big Bang Scenario in String Cosmology," *Phys. Rep.* 373, 1, Preprint: http://arxiv.org/abs/ hep-th/0207130.
Vilenkin, Alexander (1982), "Creation of universes from nothing," in *Phys. Lett.* B117, 25.
Vilenkin, Alexander (1994), "Approaches to Quantum Cosmology," in *Phys. Rev. D* 50, 2581–2594, Preprint: http://lanl.arxiv.org/abs/gr-qc/9403010v1.
Vilenkin, Alexander (2002), "Quantum cosmology and eternal inflation," in *The Future of Theoretical Physics and Cosmology*, proceedings of the conference in honor of Stephen Hawking's 60'th birthday, preprint: http://arxiv.org/abs/gr-qc/0204061.
Vilenkin, Alexander (2006), *Many Worlds in One*, Author: Alexander Vilenkin, Publisher: Hill & Wang.
Wald, Robert M. (1984). *General Relativity*, Publisher: The University of Chicago Press, pg. 215–16.

Zhang, Jinfei, Zhang, Xin, Liu, Hongya (2007) "Holographic dark energy in a cyclic universe," Authors:. Eur. Phys. J. C 52 (2007) 693–699 Preprint: http://arxiv.org/PS_cache/arxiv/pdf/0708/0708.3121v2.pdf.

Zhang, Xin (2007a), "Can black holes be torn up by phantom in cyclic cosmology?," Preprint: http://arxiv.org/PS_cache/arxiv/pdf/0708/0708.1408v1.pdf.

Zhang, Xin (2007b), "Comment on "Turnaround in Cyclic Cosmology" by Paul Frampton & Lauris Baum", preprint: http://arxiv.org/abs/0711.0667v1.

2

Why the Big Bang Singularity Does Not Help the *Kalām* Cosmological Argument for Theism

J. Brian Pitts

1 Introduction

The epistemic status of theistic belief continues to be of interest. The modern western intellectual climate being skeptical about divine revelation and the supernatural in general, theists seem to be well served if good theistic arguments from natural philosophy are available. One of the more popular theistic arguments, given current science and especially the success of Einstein's General Theory of Relativity (GTR) as applied to cosmology, employs the Big Bang singularity to show that the universe began to exist. This theistic apologetic strategy often proceeds informally, but sometimes by explicit use of the formal *Kalām* cosmological argument. Whatever the details, this type of argument has commended itself to not a few reputable physicists (as discussed in Jaki [1980]) and other scientists and philosophers, as well as at least one Pope, Pius XII (Pius XII [1952]; McMullin [1981a]). In the 1960s, Engelbert Schücking announced that '[a]lthough some cosmologists might still unconsciously project theological pictures into their science, we have been able to scare most of the ministers out of cosmology by a straightforward application of tensor calculus' (Schücking [1967], p. 218). Schücking notwithstanding, significant and visible efforts continue to be made to support theism *via* the Big Bang singularity. Recent endorsements include the human genome expert Francis Collins (Collins [2007], pp. 64–7) and the new deistic Antony Flew (Flew and Varghese [2007], pp. 135–8). Evidently, some other atheists besides Flew have agreed that Big Bang cosmology supports theism, and so have doubted Big Bang cosmology:

> Perhaps the best argument in favour of the thesis that the Big Bang supports theism is the obvious unease with which it is greeted by some atheist physicists. At times this has led to scientific ideas . . . being advanced with a tenacity which so exceeds their intrinsic worth that one can only suspect the operation of psychological forces lying very much deeper than the usual academic desire of a theorist to support his/her theory. (Isham [1997], p. 378).

While some theists and some atheists hold that the singularity supports theism, both atheists and theists (Drees [1990]; Küng [2007]) have criticized the argument as well. However, atheistic critics not infrequently mix with their good points various unhelpful moves such as denying *ex nihilo, nihil fit* and thus perhaps ceding the rational high ground, introducing premises that strike the theistic apologist as question-begging, or writing in a tone suggesting that rejection of the theistic conclusion plays an undue role in motivating the rejection of the argument. Comparable remarks could be made about some criticisms made by some liberal theists: their *a priori* hostility toward special or detectable divine action in the world might give the impression of bias against arguing from the singularity to theism, and thus blunt the force of their good criticisms.

Judging by the lack of progress in some of the literature, the result is a stalemate (Craig and Smith [1993]; Redhead [1996]). Quentin Smith, who has played a role in critiquing the singularity argument for theism, has a different but also negative evaluation of the overall response to the theistic argument from the Big Bang singularity. Smith ([1992]) held: 'The response of atheists and agnostics . . . has been comparatively weak, indeed, almost invisible'.

Thus there is room, I take it, for a critique of the argument from the Big Bang singularity to theism that strives to convince even the Big Bang singularity argument's proponents, not just the indifferent and the already opposed. If an argument is bad, there is value in showing its badness even to its proponents, if possible. Such a critique will require a brief discussion of the doctrine of creation. It will also involve some concessions to the Big Bang theistic apologist that might otherwise seem overly generous; such concessions will make evident how weak and non-question-begging the premises used in the critique are. The work will have a bit of the flavor of a review article.

The critique presented here also has some novel elements, which should be of direct interest to nontheists and fence-sitters as well. One of these elements is a *reductio ad absurdum* strategy that leads to the conclusion that the Big Bang singularity implies a divine Creator only if stellar gravitational collapse to a black hole implies a theistic Destroyer. Apart from some differences that do not ultimately matter, the scenarios are the same. A second novelty is a strengthening of the conventionality worries by appeal to theories of gravity other than GTR. The Bach–Weyl theory of gravity shows that a physical theory need not even define the length of a curve. In scalar–tensor theories or other theories with multiple metrics, age might be radically ambiguous. Thus, only a topological notion of 'beginning' in terms of a first moment is available if the *Kalām* argument is intended to yield a necessarily nonvacuous necessary truth. But a first moment is easy to avoid, as will appear below. I will also suggest that some Leibnizian criticisms of the argument from the Big Bang to theism are less persuasive than they are often taken to be.

It should be emphasized that the point at issue is not primarily the soundness of the *Kalām* cosmological argument, but rather its dialectical effectiveness. A theist might think that the *Kalām* argument is sound and accept the critique made here, by holding that the key controversial premise of the *Kalām* argument ought to be accepted not on the grounds of Big Bang cosmology, but only on the basis of divine revelation, which might or might not take the form of an argument of some kind. Theists nowadays increasingly argue that belief in God can be warranted even in the absence of argument

(Plantinga [2000]), roughly because God has designed the human mind such that when functioning properly in certain contexts, it forms belief in God. Such (indirectly) divinely caused belief formation is a suitably reliable process to produce knowledge, not just true belief by a happy accident. Thus, Plantinga claims, theistic belief might well be warranted if true; one cannot show that theism isn't knowable apart from showing it to be false (Plantinga [2000]). Clearly, this Reformed epistemology is a form of externalism—the epistemological view that the evidence, warrant, or justification for a belief counting as knowledge need *not* be available to the knower—in contrast to the internalism dominant in the tradition of Descartes and Locke. Whatever the value of Reformed epistemology, the significance of its partial rebellion against Locke's anti-Enthusiast measures (Mavrodes [1989]), or the intricacies of epistemologists' debate over internalism *versus* externalism, the task in this paper is the evaluation of a theistic *argument* of the cosmological sort, in both the philosopher's and the physicist's senses of 'cosmology'. The medieval debate about Aristotle's doctrine of the eternity of the world provides a very helpful precedent.

2 The doctrine of creation and its warrant

Historically, the doctrine of creation has been basically common property among Judaism, Christianity, and Islam. While there have been some differing views about the doctrine, the differences have little to do with the differences between these religions. A consensus arose that, *pace* Plato's *Timaeus,* creation was *ex nihilo*: God brought (or timelessly brings) all other things into existence and does so without using preexisting materials. Furthermore, God creates the world voluntarily, as a result of willing rather than as an unwilled necessary emanation. Besides the initial creation *ex nihilo* 'in the beginning', there is some sort of ongoing sustenance, preservation, continuing creation or the like, by which God keeps the world in existence and without which it would cease to exist (Quinn [1993]).

Given creation *ex nihilo,* a secondary question is whether God created the world in the finite past or eternally. We may speak of 'creation in time' if the universe is finitely old and had a first moment of existence. (If the universe is finitely old but had no first moment, then it is unclear whether creation in time or eternal creation is a more appropriate term, especially on account of conventionality worries to be discussed below.) At times, some have denied the initial creation in favor of God's eternally creating the universe. In medieval times, this denial occurred with not-so-orthodox Islamic philosophers such as Avicenna and Averroes and with the very Aristotelian Averroists in western Christendom. This 'eternal creation' view was defeated by orthodoxy defending creation in time. In western Christendom, this defeat took institutional form in the Fourth Lateran Council in 1215 and the condemnations by Bishop Tempier of Paris in 1277.

Advocates of creation in time have disagreed whether the finite age of the universe could be known by reason or only by faith (Thijssen [1998]). In response to Aristotelian philosophical arguments for an eternal universe, Moses Maimonides

and Thomas Aquinas both held such arguments to be inconclusive. Both also held the philosophical arguments of their days in favor of a finite age to be inconclusive. Thus the issue was left to be resolved by Scripture, which favored finite age (Hyman and Walsh [1973]). Bonaventure, by contrast, took the finite age of the universe to be demonstrable philosophically. This distinction is closely analogous to the issue addressed in this paper. Those who argue for theism from the Big Bang singularity follow Bonaventure in spirit. I will argue that theists who affirm creation in time ought to hold, with Maimonides and Aquinas in spirit, that creation in time is known by faith rather than by natural philosophy—or at least, not from the singularity in Big Bang cosmology.

Apart from the thirteenth-century Aristotelian controversy over the eternity of the world, there seems to have been little controversy in Christian circles about creation in time until modern theology arose. As Langdon Gilkey writes, '[u]ntil the nineteenth century, almost every Christian thinker accepted this dual implication [involving a temporal origin as well as metaphysical dependence] of the doctrine of creation: God long ago brought the world into existence out of nothing at a first moment of the time series. To all of them it seemed the clear message of the first verse of Genesis' (Gilkey [1959], p. 311). In modern theology, the initial creation was downplayed by Friedrich Schleiermacher as not necessitated by his controlling principle of the feeling of absolute dependence (Russell [1996b]). Moreover, he took the doctrine of the initial creation event to depend on the *Genesis* creation account, which he took to be the product of a mythological time (Copan and Craig [2004], pp. 150, 151). Quite a few modern scholars have felt free to discard or hold lightly the initial creation event, so eternal creation is now regarded as attractive in some circles (Gilkey [1959]; Polkinghorne [1994]; Russell [1996a]; Helm [1997]). According to John Polkinghorne, '[t]heology is concerned with ontological origin and not with temporal beginning. The idea of creation has no special stake in a datable start to the universe' (Polkinghorne [1994]). Paul Helm has agreed (Helm [1997]). On the other hand, Phil Quinn offers an account in which creation implies a first moment; thus if an object has no first moment, God doesn't create it, though he conserves it in existence (Quinn [1993]). The question has also arisen whether Genesis 1:1 actually asserts a beginning or not, but the traditional interpretation still certainly has defenders (Copan and Craig [2004]). It should go without saying that the interpretation of a text is a distinct issue from the correctness of its assertions.

If a universe is created, is it eternal, finitely old but lacking a first moment, or finitely old with a first moment? Presumably, time is isomorphic to the real numbers, the integers, or a connected subset (having no holes in the middle, so to speak) of either. The universe's having a first moment entails it being finitely old (unless one introduces a point at infinity, which seems purely formal), but the converse entailment might fail. If time is continuous, then the world's being finitely old does not entail it having a first moment. At least *prima facie* it seems that a universe with a first moment is the sort most confirmatory of theism (McMullin [1981b]), while an eternal universe is the least helpful. Regarding the two extreme cases, Aquinas held something along these lines, as Ernan McMullin discusses (McMullin [1981a], p. 39):

> If the universe began at a point of time, would this give stronger support to the claim that a Creator is needed than if the universe always existed? Aquinas argued that in a sense it would, even though he was insistent that a universe which had always existed *would* equally need a Creator to sustain it. But creation in time rather than from eternity makes the work of God's power more evident, Aquinas says, because an agent displays the more power in acting, the more removed from act is the potency acted upon. And in creation in time there is no potency of any kind to work on. This of itself immediately shows the infinity of power required to summon a universe into act.

It seems that a first moment conveys three advantages on a doctrine of creation over mere finite age.

First, it is more obvious that a finitely old universe with a first moment requires an external cause than that a finitely old universe in which every moment is preceded by an earlier one does. In the latter case, one might be tempted to think that the present is fully explained by the past within the history of the universe, so nothing external is required (Grünbaum [1989]; Earman [1995]; Smith [2000]).

Second, having a first moment is a topological notion, not a metrical one, and so escapes conventionality worries (Misner [1969]; Agnese and Wataghin [1971]; McMullin [1981a]; Grünbaum [1989]; Levy-Leblond [1990]) about temporal remetrization of a finite past to an infinite one. Conventionalist questions about the significance of the difference between finite *versus* infinite seem to be due to E. A. Milne (Kragh [2004], p. 209; McMullin [1981a]). Conventionality becomes an especially serious worry in physical theories containing multiple metrics. Scalar–tensor theories are perhaps the best known locus for the question 'Which metric is the physical metric?' (Magnano and Sokolowski [1994]; Weinstein [1996]; Kaloper and Olive [1998]; Santiago and Silbergleit [2000]). While this question seems not to need an answer for most purposes (such as those involving only the field equations), questions of singularities, boundary conditions, positive energy, and quantization give that question a bit more urgency (Kaloper and Olive [1998]; Santiago and Silbergleit [2000]; Faraoni and Nadeau [2007]; Catena *et al.* [2007]; Sotiriou *et al.* [2008]). Theories with multiple metrics might have different types of matter coupling to gravity in different ways; then perhaps one metric might yield finite age, but another infinite age, in which case there seems to be no answer to the question 'how old is the universe?', even if the options at hand are merely 'finite' and 'infinite'. It turns out that the actual universe probably does not behave in accord with a scalar–tensor theory, given the empirical confirmation of the various principles of equivalence in gravity (Will [1993]), which make it difficult for theories empirically distinguishable from GTR in weak or moderate gravitational fields to be empirically viable. But a doctrine of creation needs to be modally rich enough to accommodate the possibility of God's creating worlds with physical laws without a unique or preferred metrical structure to license an answer of 'finite' or of 'infinite' age.

Third, a first moment, being topological rather than metrical, is well defined even in the absence of a metric. Bach–Weyl conformal gravity (Fiedler and Schimming [1980]; Schimming and Schmidt [1990]; Kazanas and Mannheim [1991]; Dzhunushaliev and

Schmidt [2000]), in the absence of matter or with some types of matter, employs only the conformal part $\hat{g}_{\mu\nu}$ of a metric, a tensor density with weight $-\frac{1}{2}$ (in four space-time dimensions) and determinant −1 (Anderson [1967]). Given a metric $g_{\mu\nu}$, one can obtain its conformal part using

$$\breve{g}_{\mu\nu} = g_{\mu\nu}\left[-det\left(g_{\alpha\beta}\right)\right]^{-\frac{1}{4}} \tag{1}$$

(in four space-time dimensions). Not having a metric, the Bach–Weyl theory takes $\hat{g}_{\mu\nu}$ as primitive and does not assign lengths to curves.[1] While light-like (null) geodesics are well defined in the Bach–Weyl theory, their affine parametrization is not (Wald [1984], p. 446), so even light-like (null) geodesics provide no help in defining finite age. By contrast, Copan and Craig simply take for granted the existence, and perhaps the uniqueness, of the metric for timelike curves in setting up the *Kalām* argument (Copan and Craig [2004], p. 199) (here presented a bit differently from the version quoted below):

> To assess the truth of the premise [that the temporal series of past, physical events is not beginningless], it will be helpful to define some terms . . . . In order that all the events comprising the temporal series of past events be of equal duration, we arbitrarily stipulate some event as our standard . . . . By a 'beginning,' one means a first standard event. It is therefore not relevant whether the temporal series had a beginning point (a first temporal instant).

Copan and Craig perhaps take their *Kalām* argument to express a necessary truth that applies nonvacuously to all possible physical theories. However, their criterion for a beginning is meaningless for the Bach–Weyl theory because 'equal duration' is meaningless in that theory. The Copan–Craig criterion might be ambiguous for scalar–tensor theories, which is also somewhat worrisome. Given that neither existence nor uniqueness of a metric (for timelike curves) holds necessarily, the natural move is to adopt a topological rather than a metrical notion of beginning. Thus a first moment *is* the point that needs to be addressed. However, to adopt a first moment as the criterion for a beginning is to admit defeat, as far as arguing from the singularity to theism is concerned, because the two plausible moves relative to contemporary cosmology (*viz.,* taking space-time to contain only points 'after' the singularity or invoking some perhaps presently unknown theory that resolves the singularity, and so permits an infinite past) lack a first moment. Given the demonstrable contingency of the existence and uniqueness of a metric, and the contingent relationship between the tensor $g_{\mu\nu}$ and physical lengths (which depends on how matter couples to gravity (Brown [2005]), it is unclear why finite metrical age is so important even for GTR.

As it turns out, the Fourth Lateran Council in 1215 made it a doctrine of faith, at least for Roman Catholics, that the world had a temporal beginning, and even escapes the conventionalist worry by positing a first moment (McMullin [1981a], pp. 29, 54, 55) (cf. Grant [1994], p. 83; Leith [1982], p. 57). The council's definition apparently

is topological rather than metrical in character, so its intelligibility withstands the mathematical advances of the last 700–800 years. Whatever one makes of the contemporary relevance of the canons of the Fourth Lateran Council as a deliverance of faith—only traditional Roman Catholics will feel bound to accept the council's teaching, while its stipulation that the angels were created at the same moment (Leith [1982], p. 57) poses an additional burden for belief—this conciliar conclusion about a first moment of the physical universe is clearly not a deliverance of reason.

As McMullin and Aquinas observe, creation in time does have some theological advantages over eternal creation, even apart from adherence to the relevant part of the Genesis account. On the other hand, the epistemology of creation, at least for the author of the letter to the Hebrews in the New Testament, might favor Aquinas rather than Bonaventure. A key passage is Hebrews 11:3: 'By faith we understand that the universe was created by the word of God, so that what is seen was not made out of things that are visible' (ESV). In light of this passage (similar in other translations, though not in the RSV), perhaps Christians should expect that philosophical or scientific demonstrations for creation in time will fail. In any case, current demonstrations from the Big Bang singularity fail.

3 Cardinality and sizes of infinity

A few brief comments on Craig and Copan's *a priori* arguments for a finite past will be useful. Regarding the supposed impossibility of traversing an actual infinity (Craig [1979]; Copan and Craig [2004]), one might object to the treatment of infinity (Morriston [2003]; Redhead [1996]). Concerning the possibility of the existence of an actual infinity, there is no evident absurdity in Hilbert's Hotel, which has as many rooms as there are positive integers. It is also difficult to regard as omnipotent a God who could not create Hilbert's Hotel. If the occurrence of a beginning of time could be known *a priori,* then arguments from physical cosmology would be largely redundant. For the sake of argument (and other reasons as well), I therefore assume that an infinite past is metaphysically possible.

It appears that Craig's objections to the possibility of actual infinities, such as Hilbert's Hotel, are due in no small part to the widespread claim that the concept of size is fully captured by the notion of cardinality. But one might well accept the possibility of actual infinities, such as Hilbert's Hotel, while doubting that cardinality exhausts the notion of sameness of size or counting. Indeed, contemporary physics implicitly denies that cardinality exhausts the notion of sameness of size or counting for infinite sets: physicists routinely count the number of degrees of freedom in a field theory as two at each spatial point (written as $2\infty^3$ in three spatial dimensions), or three at each spatial point (written as $3\infty^3$), or the like (Sundermeyer [1982]; Henneaux and Teitelboim [1992]). Whereas Maxwell's electromagnetic theory has $2\infty^3$ degrees of freedom, Proca's electromagnetism with massive photons (to borrow a quantum phrase), a theory (or family thereof) worthy of consideration in discussions of underdetermination and empirical equivalence, has $3\infty^3$ degrees of freedom. Physicists never feel bound to

perform, or even entertain, such Cantorian manipulations as $2\infty^3 = \infty$, $3\infty^3 = \infty$, and therefore $2\infty^3 = 3\infty^3$, so that Maxwell's and Proca's theories have the same number of degrees of freedom. Such manipulations would reduce some important reasoning in field theory and constrained dynamics to nonsense. The lesson seems to be that physical theories involve continuity properties of sets from which cardinality abstracts, and that it seems reasonable to call the results 'counting' nonetheless. Evidently, cardinality does not exhaust the useful notions of 'same size' or counting in all contexts. A related point pertains to the mathematics of Fourier analysis of arbitrary functions in terms of sinusoidal functions: for finite intervals of the real line, one can represent functions using a Fourier series with countably infinitely many coefficients, whereas for the whole real line one represents functions using a Fourier integral with uncountably many coefficients (Jackson [1975]), though the finite interval and the whole real line have the same cardinality. These important parts of applied mathematics do not obviously have a straightforward connection to cardinality. It seems inadvisable, then, to say that cardinality simply *is* the notion of size or counting for infinite sets (though its technical utility in some contexts is undeniable). Thus one can perhaps sympathize with some of Craig's mathematical motivations for *a priori* objections to actually infinite collections. However, no good argument against their possibility results.

4 Modern cosmology and creation

Modern physical cosmology is a rather effective framework for unifying and explaining a wide variety of astronomical observations in a framework provided by well-confirmed physical laws that hold here and now. Roughly speaking, one assumes a Robertson–Walker space-time metric satisfying Einstein's gravitational field equations (Wald [1984]). This metric is spatially homogeneous and isotropic, meaning that every point in space is alike at a given moment of time, and so is every direction. Clearly this is an idealization. Data from the present, especially the fact that luminous objects' redshifts are larger for more distant objects, indicate that the universe is expanding (*modulo* conventional redescription, which might allow depiction of all objects as shrinking). The dynamics of GTR, with standard kinds of matter, let one extrapolate back to an earlier hot dense phase, during which time plausibly the observed cosmic abundances of light elements were produced. During the hot dense phase, the material contents of the universe would be radiation or matter closely approximating radiation in its equation of state relating pressure and density. Given the form of the stress–energy–momentum tensor for radiation, GTR predicts (or retrodicts) that the universe would expand roughly as $a(t) \sim \sqrt{t}$, where a is the scale factor and t measures physical proper time, the 0 corresponding mathematically to the Big Bang singularity. This behavior is exact for the (idealized) spatially flat case, and holds as an increasingly good approximation as $t \to 0$ in the positively or negatively spatially curved cases (Wald [1984], p. 98).

One theoretical problem that frequently has been neglected is the 'averaging problem', the need to average Einstein's equations over cosmic distances in order

to find equations for the cosmic parameters (Ellis [1984]; Zalaletdinov [1992]; Buchert and Carfora [2002]; Coley and Pelavas [2006]). The analogous procedure for electromagnetism in a medium is well known and comparatively simple due to the linearity of Maxwell's equations (Jackson [1975]; Mars and Zalaletdinov [1997]). For Einstein's equations, the dynamics of the average bears a complicated relation to the average of the dynamics, so to speak, due to nonlinearities. Experimentally, since the late 1990s it has appeared that the cosmic expansion is accelerating. However, some contend that paying attention to the averaging problem might help to resolve the apparent phenomenon of accelerating expansion, for which 'dark energy' has been posited (Buchert [2008]) as an explanation or at least a name. It has also been argued that part of the persuasiveness of the acceleration of the cosmic expansion is an artifact of conventional choices of statistical variables (Cattoen and Visser [unpublished]), or that a deeper failure of homogeneity might require consideration (Ellis [2008]). Little turns on these issues for my purposes, because using the most current version of Big Bang cosmology with accelerating expansion does nothing to strengthen the Big Bang singularity argument for creation and time, and hence theism. The same could be said for inflation, which John Earman and Jesus Mosterin have examined critically (Earman and Mosterin [1999]). I will freely ignore these more recent developments in favor of the old hot Big Bang cosmology.

Given the observed cosmic expansion, mathematical extrapolation still further into the past—a bold move that might have little warrant, depending on how far one extrapolates—implies that the space-time metric was singular roughly 14,000,000,000 years ago: there was a state of infinite curvature and density, through which Einstein's equations allow no further retrodiction. It should be emphasized that extrapolation over small periods of time near $t = 0$ corresponds to extrapolation through an extraordinary range of energies, into realms far beyond the reach of particle accelerator experiments. Contrary to some earlier hopes (Earman [1995], p. 205) still present in the 1960s, general relativistic singularity theorems show the singular behavior to be generic, not an artifact of the high symmetry assumed in homogeneous isotropic or spherically symmetric models as it is in Newtonian gravity (Wald [1984]). If homogeneity is assumed (as it usually is, at least for the prototypical models such as Robertson–Walker), then the singularity occurs everywhere throughout all space, so it has seemed natural to speak of an origin of space (or space-time) at the singularity. However, homogeneity on scales beyond the horizon is an assumption or convention (Bondi [1947]; Layzer [1954]; McCrea [1955]; Callan *et al.* [1965]; Klein [1971]; Feynman *et al.* [1995]; Smoller and Temple [2003]; Adler *et al.* [unpublished]), not an empirical fact. Indeed, even the global topology of space-time is subject to worries regarding conventionality (Glymour [1973]; Malament [1977]). Empirically we have no (direct) access to regions more distant than some billions of light years (at least apart from quantum mechanics), due to relativistic causality constraints. As a result, claims that homogeneity holds out to, say, a trillion light years, or 10^{30} light years, are not observationally well grounded (Earman [1995], p. 125; McCabe [2004]). (Of course, inflationary cosmology could complicate matters.) I emphasize the spatially local rather than global nature of

current cosmological knowledge, and hence of the singularity, in anticipation of discussing the analogy between Big Bang cosmology and the time reversal of stellar gravitational collapse. If one wishes to think of Big Bang cosmology as involving, say, a sufficiently vast expanding ball of matter surrounded by empty space, the data supporting hot Big Bang cosmology provide no obstacle; of course, the Robertson–Walker metric will not apply in the exterior vacuum region. A bit of historical perspective is useful here. Though there were exceptions, '[i]n the ideological debate of the late 19th century, finitism—the view that the world is finite in time and space—was usually associated with conservativism and Christian belief, whereas socialists and materialists adhered to the doctrine of an infinite and eternal universe' (Kragh [2004], p. 57). It is not difficult to see how finitism could serve teleological world views by reducing the probabilistic resources for unguided evolutionary processes, or how its denial could serve nonteleological world views, but experiments have not resolved the issue and perhaps cannot.

Just 'where' in the space-time metric the Big Bang singularity lies deserves some comment. Inverting a relation above, one can break the space-time metric down into a conformal factor $\hat{g}_{\mu\nu}$ and a factor of the metric's determinant, which determines volumes

$$g_{\mu\nu} = \breve{g}_{\mu\nu}\left[-\det\left(g_{\alpha\beta}\right)\right]^{\frac{1}{4}}. \tag{2}$$

Given that Big Bang cosmology is glossed as involving the expansion of the universe, and that this expansion is isotropic, one might think that the metric tensor $g_{\mu\nu}$ becomes singular in the part that determines volumes, the determinant $g =_{def} det(g_{\mu\nu})$. That cannot be the case, however, because one can achieve $g = -1$ in some neighborhood about any given point, a well-known fact that recently was seen to have deep consequences for analyzing the 'general covariance' that supposedly characterizes the conceptual innovation of GTR (Pitts [2006]; Giulini [2007]). The Big Bang singularity cannot be a property of $\hat{g}_{\mu\nu}$, because $\hat{g}_{\mu\nu}$ is the same as in the Minkowski flat space-time of Special Relativity, on account of the conformal flatness of the Robertson–Walker metrics (Infeld and Schild [1945]; Tauber [1967]). For the same reason, the Big Bang singularity cannot manifest itself in bad behavior of the Weyl curvature tensor, which in fact vanishes for the Robertson–Walker metrics due to their conformal flatness.[2] It follows that the Big Bang singularity manifests itself in the misbehavior of the relation between $\hat{g}_{\mu\nu}$ and g, which also manifests itself in the Ricci (or equivalently, the Einstein) curvature tensor.

While a number of people noted above have regarded the Big Bang singularity as potent evidence for creation in time and hence for theism, perhaps the most visible sustained defenses have come from Craig (Craig [1979]; Craig and Smith [1993]; Copan and Craig [2004]) and from astrophysicist–apologist Hugh Ross (Ross [1991], [2001]). While Ross's works generally are not aimed at the academy, they receive endorsements by some reputable physicists and astronomers such as Donald Page and Allan Sandage, generously cite technical papers and have some semi-popular influence. As Ross puts the issue,

> [i]n Hawking's words, time itself must have a beginning. [footnote omitted] Proof of the beginning of time may rank as the most theologically significant theorem of all time, assuming validity of the general theory of relativity (Ross [2001], p. 102).

The conditionalizing on GTR soon disappears, however, with a ringing endorsement of that theory:

> Today it can be said that no theory of physics has ever been tested in so many different contexts and so rigorously as general relativity. The fact that general relativity has withstood all these tests so remarkably well implies that no basis at all remains for doubting the conclusions of the space-time theorem (Ross [2001], p. 107).

By contrast, William Lane Craig argues more carefully and concludes more modestly. Craig formulates the *Kalām* cosmological argument along these lines.

(1) Everything that begins to exist has a cause of its existence.
(2) The universe began to exist.
(3) Therefore the universe has a cause of its existence (Craig [1979], p. 63).

This argument is valid, at least if it is unproblematic to treat the universe as a thing, rather than a collection of things; I assume that there is no problem here. The first premise seems true (apart from worries about virtual particles in quantum field theory, perhaps), at least on some readings of 'beginning', including the one for which I argued above, namely, a first moment—though that is not Craig's reading. The truth of the second premise, or rather, the source of warrant for the second premise if it is true, is the key question. Many theists will affirm the second premise and regard the *Kalām* argument as sound, but what is the reason for affirming the second premise? Clearly, one will not persuade the nontheist to accept theism (or strengthen the theist's faith with scientific support), as the argument is intended to do, if the warrant for the second premise comes wholly from ostensible divine revelation. If one accepts, say, the Bible as divinely inspired, then one has already accepted theism and much else besides.

5 Tolerance or intolerance toward singularities?

The question therefore arises whether to take the singularity seriously as a feature of the real world, or to dismiss it as an artifact of incomplete physical understanding. As one sees all the time in papers on quantum gravity, most people who work on quantum gravity take for granted that the Big Bang singularity is an artifact of incomplete physical understanding and expect or hope that uniting gravity with quantum mechanics in some kind of quantum gravity will resolve the singularity into some well-defined situation that admits extrapolation to still earlier times, *ad infinitum*. Jayant Narlikar has persuasively deployed this point as a critique of the argument from the singularity to theism (Narlikar [1992]). Earman, by contrast, defends a less widely held view about

singularities, namely, that we should display 'Tolerance for Spacetime Singularities' (Earman [1996]) and try to learn from them.

Tolerating singularities and trying to learn from them, as Earman does, is an attitude that should commend itself only to GTR-exceptionalists, those who emphasize the differences between GTR and the other forces over the similarities between them. If one thinks that gravity as portrayed by GTR is importantly like other forces, then gravitational singularities are not appreciably more interesting than the singular electric field of a point charge, which simply needs to be resolved by a better theory, such as quantum electrodynamics. Thus 'complete gravitational collapse [in that case is u]nimportant or at most peripheral' (Misner *et al.* [1973], p. 437). But if gravity differs importantly from the other forces, one might conclude that 'complete gravitational collapse [is c]entral to understanding the nature of matter and the universe' (Misner *et al.* [1973], p. 437). Then, however, one has strong technical reasons to doubt that singularities exist as part of space-time.[3] Within the Robertson–Walker cosmological space-time for $t > 0$ (which is to say, always), one can explain each moment in terms of an earlier one (Grünbaum [1989]; Earman [1995]; Smith [2000]). Thus there is no beginning required and premise 2 might be false, as far as physics can tell. Those who strive to 'take the lessons of relativity seriously', as GTR-exceptionalists do, might also have reason to doubt the evolutionary space-evolving-over-time picture in favor of some inherently four-dimensional picture. If one rejects the demand for such evolutionary explanations as a hangover from Newtonian physics, then again the space-time for $t > 0$ seems sufficiently self-explanatory that the singularity gives no reason to infer a Creator.

In contrast to GTR-exceptionalism, one might take the view, more common among particle physicists but not well known among philosophers, that Einstein's equations merely describe a self-interacting massless spin 2 field, much as Maxwell's equations describe a spin 1 field and the Yang–Mills equations describe a set of self-interacting spin 1 fields (Fierz and Pauli [1939]; Rosen [1940]; Papapetrou [1948]; Gupta [1954]; Kraichnan [1955]; Feynman *et al.* [1995]; Ogievetsky and Polubarinov [1965]; Weinberg [1965]; Sexl [1967]; Deser [1970]; Weinberg [1972]; van Nieuwenhuizen [1973]; Veltman [1981]; Fang and Fronsdal [1979]; Pitts and Schieve [2001]; Boulanger and Esole [2002]); the distinctive features of GTR are seen as incidental (though important) technical consequences of gravity having spin 2 rather than spin 1. While probably all who tolerate singularities are GTR-exceptionalists, many or most GTR-exceptionalists, such as those who work on canonical quantum gravity and loop quantum gravity, do not tolerate singularities.

It appears, then, that whether one is tolerant or intolerant toward singularities, it turns out that there is no first moment (unless one is installed by hand), because every moment is preceded by earlier moments. Thus, in the relevant sense for the *Kalām* argument to be valid and to make sense for a sufficiently broad collection of physical theories, there is no beginning implied by physics, and so premise 2 might be false, as far as physics can show. In order for the Big Bang singularity to provide a good theistic argument, the singularity must be well enough behaved to be a real and intelligible part of space-time, and badly enough behaved that it cannot have a past. Satisfying both

conditions seems difficult and unlikely to be achieved. Moreover, there are various reasons, some quite good, for not tolerating singularities, which it will be worthwhile to explore.

6 Leibniz against incompetent watchmaker?

One might think that Leibniz has provided the prototype for a good critique of the singularity argument for theism. In the Leibniz–Clarke correspondence (Alexander [1956], pp. 11, 12), Leibniz famously argued against Isaac Newton and Samuel Clarke that God would not create the physical universe in such a way that it would break down and require repair from time to time. Leibniz took Newton's views to have just such a consequence, so if the world is analogous to a watch, then Newton's God is an incompetent watchmaker because Newton's God was required to perform miracles on occasion to restore the solar system to working order. Whereas one sometimes encounters biological dysteleology arguments, this is a physical dysteleology argument. Whether Newton and Clarke deserved this criticism need not concern us. Acceptance of Leibniz's analogy between the physical world and a watch did not require a rejection of miracles, for Leibniz accepted miracles of grace, while rejecting miracles posited to fix nature due to poor design (Alexander [1956], p. 12). If the physical world is like a watch, then it ought to be able to run forever without breaking down. But the singularity theorems of GTR show that it cannot 'run forever without breaking down'. (There is no obvious analogy to the repair of the watch.) Thus the singularity theorems arguably show that GTR demonstrates its own inadequacy, one might conclude. Because God would not build the world so incompetently, it follows that GTR is not the correct theory for describing gravitational collapse; the true theory would not yield singularities. But GTR is time-reversal invariant, and the Big Bang singularity is simply the time reversal of a specific model of gravitational collapse of a star, which uses a Robertson–Walker metric for the stellar interior (Misner *et al.* [1973], pp. 846–59). Thus the Big Bang singularity would be eliminated along with the singularities of gravitational collapse, or so the argument might go.

This sort of argument continues to be used today, often without the explicit theism, by theoretical physicists. Thus Abhay Ashtekar, one of the dominant figures in contemporary work on quantum gravity, opened a recent review of the field with the following motivation:

> *Big-Bang and other singularities:* It is widely believed that the prediction of a singularity, such as the big-bang of classical GTR, is primarily a signal that the theory has been pushed beyond the domain of its validity. A key question to any quantum gravity theory, then, is: What replaces the big-bang? . . . (Ashtekar [unpublished]).

One can find similar sentiments elsewhere very easily among workers in quantum gravity.

Leibniz might or might not have been entitled to his 'God wouldn't do it that way' premise. He lived in a still Christian age and took himself to know many things by divine revelation; many contemporary physicists cannot say the same. If one is doing natural rather than revealed theology, the premise about what God would or wouldn't do seems difficult to justify. Thus Elliott Sober has recently warned against this sort of *a priori* theological claim about what God would or ought to do (Sober [2003], [2007]). Whereas Stephen Jay Gould seemed to think that he knew that God (if he had existed) wouldn't have made the panda's thumb as it is, Sober denies that Gould knew any such thing about the dispositions of the deity.

The Leibnizian intuition that the world would not break down after a finite time has a certain appeal for physicists, but reflection suggests that it might be difficult for the argument to get traction with those not already disposed to accept it, except perhaps for theistic rationalists such as Leibniz. That this Leibnizian intuition (perhaps stripped of its theistic justification) in fact appeals to physicists may reflect the fact, discussed in theories of science by Thomas Kuhn, Imre Lakatos, and others (Laudan *et al.* [1986]), that ordinary scientific progress depends upon a certain faith that things are on the right track and that progress is possible; anomalies that could serve as counterevidence are in fact typically not regarded as threatening the paradigm or hard core of the research programme, but rather as opportunities to display its excellence, sooner or later, by triumphing over difficulties, while causing no alarm in the mean time. The Leibnizian intuition might well be truth-conducive, but it is difficult to argue for that conclusion. This phenomenon bears some resemblance to the sort of stalemate that can result in the scientific realism–antirealism debate (Kukla [1998]). However, the sort of intellectual pessimism or strong empiricism that rejects the Leibnizian intuition, if accepted, might undermine the explanatory drive that motivates accepting premise 1 of the *Kalām* argument.

7 Induction from earlier theories' breakdown?

The physics of a century ago provides some material for an inductive argument that quantum mechanics will fix the problems in Einstein's classical theory of gravity. The story of blackbody radiation and the development of quantum theory is complicated (Kuhn [1978]), more so than the view given in modern physics textbooks, but the following selection should suffice. A century ago, there were good classical theoretical arguments for the Rayleigh–Jeans law for blackbody radiation, according to which the energy density for radiation at a given frequency increased with frequency. (The Rayleigh–Jeans law was also demonstrably false empirically, but in some ways that is an irrelevant accident for present purposes. Lakatos has already extolled the virtues of rationally reconstructed history.) But such a radiation law has to be wrong, because integrating over all frequencies (up to $+\infty$) implies that a blackbody radiates *infinite* power. Ordinary objects, especially black ones, approximate blackbodies, so they would radiate away their energy immediately in a blinding flash, contrary to experience. This difficulty has come to be known as the 'ultraviolet catastrophe'.

The answer of Planck and others to the threatening inference of ultraviolet catastrophe, history shows, helped to lead to quantum mechanics. New theoretical foundations were brought in that yielded an exponentially decaying factor to counteract the Rayleigh–Jeans power law growth, and thus give a convergent integral up to infinite frequency. (The exponentially decaying factor preceded widespread worries about the ultraviolet catastrophe, but previously it was motivated on more empirical grounds.) From a logical point of view, the ultraviolet catastrophe was a *reductio ad absurdum* of the classical physics underlying the Rayleigh–Jeans radiation law. The solution was new physics of a quantum kind, which averted the catastrophic infinity; no miracle was required. Using this case and others, one might argue inductively that just as quantum mechanics resolved these problems, so it will resolve the singularity problems of GTR.

Earman insightfully discusses a similar argument to the effect that GTR contains the seeds of its own destruction (Earman [1996]), citing Kip Thorne's invocation of the Rutherford atom. However, as Earman notes, the Rutherford atom with classical electromagnetism was empirically inadequate, something that cannot be said at present for GTR's prediction of gravitational collapse. A theory that predicts disaster here and now clearly needs to be changed, if disaster is not observed. However, GTR fits the data rather well. What of its ultimate mathematical breakdown in singularities? Earman observes that if GTR is taken at face value, then there just is not some region 'beyond' the singularity that the theory fails to describe. It would seem that the supposed inductive argument relies on a metatheoretic criterion besides empirical adequacy to determine that GTR breaks down, one that strongly resembles the Leibnizian intuition previously discussed. Thus '[t]he analogy with the Rutherford atom is not apt' (Earman [1996]). So this inductive argument will not persuade those who are not persuaded by the Leibnizian intuition. But there are more compelling reasons for rejecting the argument from the Big Bang to theism.

8 Stellar collapse implies theistic destroyer

One might wonder why theological significance should be ascribed to the Big Bang singularity, but not to other physical singularities. This worry takes its most acute form when one considers the similarity of the Big Bang cosmology to the time reverse of the gravitational collapse of a star to a black hole with a central singularity. Assuming homogeneous matter distributions, both Big Bang cosmology and stellar gravitational collapse use the Robertson–Walker space-time metric in the matter-filled region (Misner *et al.* [1973], pp. 846–59). If one chooses a Big Bang model with a bounded matter distribution, as one certainly may (Bondi [1947]; Layzer [1954]; McCrea [1955]; Callan *et al.* [1965]; Klein [1971]; Feynman *et al.* [1995]; Smoller and Temple [2003]; Adler *et al.* [unpublished]) (though typically one doesn't, partly for convenience), then both the Big Bang and the gravitational collapse have a Robertson–Walker interior matched to a Schwarzschild exterior. The main differences are the direction of time and the distance scale. The distance scale does not seem important for present purposes. If the Big Bang strongly indicates that there exists a God who created the universe,

do formally similar time-reversed events such as the gravitational collapse of stars to form black holes with singularities imply that there exists a God who supernaturally destroys (annihilates, ceases to uphold) the interiors of stars?

That God is the Destroyer of collapsed stars is a rather surprising conclusion. While Christianity portrays God as creating, sustaining, and redeeming the world, the idea of God's absolutely annihilating either the physical world as a whole or a bit here and there from time to time appears to be a novelty or certainly a rarity in Christian theology. Of course rocks, buildings, and animals perish from time to time, but their physical remains persist, which is precisely what does *not* happen in the stellar case. Orthodoxy has held generally that even those who 'perish' in hell (presumably the best candidates for ceasing to exist) in fact persist forever, body and soul, albeit in a miserable condition. Given the divine policy of upholding the universe after creating it, for God to stop upholding, say, a star that collapses to a black hole, would be a miracle. Annihilating an object (ceasing to uphold it in existence) might not be the same kind of miracle as creating one *ex nihilo,* but it is a miracle nonetheless. I would not be appreciably less astonished by witnessing the sudden annihilation of my car than by seeing the abrupt appearance *ex nihilo* of a horse. Thus, whatever asymmetry of time might exist in relation to causation, it does not help to avoid the conclusion that stellar gravitational collapse terminates in a miracle. The proponent of the Big Bang argument therefore needs to explain why the termination of stellar gravitational collapse is not a miracle, or why such a miracle is not pointless and hence absurd. It is somewhat comforting that God, on the view in question, would only annihilate regions of high density, perhaps typically surrounded by an event horizon. Thus there is plenty of warning so that we may avoid these episodes of annihilation, and the farmer who stores wheat in a silo need not fear the annihilation of his wheat. But the theoretical oddness remains. The theological distinction between ordinary and special providence, though probably never absent historically (else the distinction between miracles and ordinary events would be meaningless or hopelessly vague), received a boost in the twelfth century in the west (Grant [1996], pp. 21, 22) and even achieved confessional status (not mere general acceptance) in the Westminster Confession of Faith (1646):

> God, in His ordinary providence, maketh use of means, yet is free to work without, above, and against them, at His pleasure (Chapter 5, Section 3) (Leith [1982], p. 200).

Given this robust (though qualified) affirmation of the general integrity of natural processes, which likely was shared outside the Reformed tradition that produced the Westminster standards, the claim that God annihilates collapsed stars would have seemed odd even to pre-Enlightenment orthodox believers (if we may imagine them informed about twentieth century physics), for whom miracles were no embarrassment. If GTR were the only possible theory to describe the divine governance of the world in matters gravitational, perhaps one could manage to accept the conclusion that God

miraculously destroys collapsed stars. But given the underdetermination of theories by data and consequent avoidability of that conclusion, that inference appears to be a *reductio ad absurdum*.

9 Stacking the deck for GTR

There is a misleadingly persuasive move made by proponents of the Big Bang argument, stacking the deck in favor of GTR and thus of Big Bang cosmology against non-singular rivals. The move is generally not made explicitly and in detail, so what follows is a reconstruction of the reasoning process that would underlie any good argument in the vicinity. It is often suggested that potentially non-singular rivals to GTR are speculative, whereas GTR is well confirmed, so GTR and its retrodiction of the Big Bang singularity ought to be accepted as the default view that challengers need to overcome with better empirical results. There is a grain of truth in this claim: some or perhaps many of the theories or models actually proposed as rivals to the Big Bang in the more speculative literature do not form part of a well-tested theory that is known to reproduce the empirical successes of GTR. Thus some of these challengers might be refuted by data already in hand, should someone think to do the necessary calculations and apply the relevant empirical data to the challengers. Were the physics community's (or its members') physical knowledge closed under logical entailment, such would-be challengers would be refuted even before publication; alas, human finitude intrudes, and it sometimes takes hard work to show whether a theory is consistent with present empirical knowledge. Thus one cannot simply collect dozens of papers that discuss non-singular cosmological models and thereby conclude that there are that many live challengers to Big Bang cosmology at the moment.

However, the grain of truth in this objection sometimes conceals the grain of falsehood that it also contains. Though not entirely trivial, it is possible to construct theories that reproduce the empirical successes of GTR in all tested regimes to date, but which differ in the ultra-strong field regime relevant to Big Bang cosmology. Granting the success of weak and medium field tests of GTR involving light bending, gravitational redshifting, time delay, and the like, why think that GTR, rather than one of its (perhaps not yet unproposed) competitors that fits the data currently in hand, is the right extrapolation? For example, one knows that massive variants of GTR (relating to GTR as Proca's electromagnetism does to Maxwell's) agree with GTR in just those regimes, while disagreeing in strong gravitational fields (Visser [1998]; Babak and Grishchuk [2003]), including those near the alleged cosmic singularity. However, positive energy has yet to be proven and might be false for massive variants of GTR (Boulware and Deser [1972]; Pitts and Schieve [2007]), so Big Bang singularity apologists can still hope that the massive theories are vicious. Apologists for (or rather, from) the singularity need to hope that a great many epistemically possible theories are vicious, in fact, a hope that most likely will disappoint eventually, if it has not already.

10 Quantum gravity tends to resolve singularities

The need to reconcile gravity and quantum mechanics all but proves that there exists a consistent theory of gravity that matches GTR in some classical limit, but which differs from it in regimes when dimensional arguments suggest that quantum effects should be large. There might well be many such theories of quantum gravity. The works of Abhay Ashtekar and collaborators, such as (Ashtekar *et al.* [2006a], [2006b]), and Martin Bojowald and collaborators, such as (Bojowald [2001a], [2001b], [2002]; Bojowald and Hinterleitner [2002]; Bojowald and Kagan [2006]), provide good evidence that the Big Bang singularity does not occur in loop quantum gravity (but see Cartin and Khanna [2005]). Loop quantum gravity is a part of the modern nonperturbative canonical quantum gravity project (Pullin [2003]), which began in the 1980s when Abhay Ashtekar proposed new variables that helped to resolve long-standing problems faced when using the older metric variables.

Moreover, a quantum theory of gravity is likely to differ considerably from GTR precisely in the ultra-strong field regime of the hot dense 'early' universe, where the Big Bang arguer relies essentially on GTR! Thus Robert Wald writes in a standard graduate textbook on GTR: 'Of course, at the extreme conditions very near the big bang singularity one expects that quantum effects will become important, and the predictions of classical GTR are expected to break down' (Wald [1984], p. 100). In such a context, curvatures comparable to the inverse square of the Planck length arise, so neglected quantum terms should be large and the classical theory becomes a bad approximation. In addition to the modern canonical quantization program, one should also keep an eye on string theory as tending to resolve singularities (Gasperini and Veneziano [2003]), not to mention various classical proposals that alter the dynamics in the strong field regime.

Lawrence Sklar once asked 'Do Unborn Hypotheses Have Rights?' (Sklar [1985]). Clearly, they do in the present context. As Bas van Fraassen has noted in the context of criticizing inference to the best explanation, '[w]e can watch no contest of the theories we have so painfully struggled to formulate, with those no one has proposed. So our selection may well be the best of a bad lot' (van Fraassen [1989], p. 143). P. Kyle Stanford argues that the problem of unconceived alternatives is an even more serious problem for scientific realism than are more commonly discussed worries (Stanford [2006], [2001]). Once the rights of unborn theories are respected, the default status allegedly held by GTR and hence of Big Bang cosmology as described by GTR near the singularity disappears. In the ultra-strong field regime of the hot dense so-called 'early universe' (to use a term that presupposes the Big Bang singularity), GTR is just another speculation among many. Is one truly rationally compelled, or even rationally encouraged, to accept an infinite extrapolation from a curve that fits the data in some finite region? Surely not. This is a severe curve-fitting problem. Sophisticated defenders of scientific realism now admit that different parts of a scientific theory are supported to different degrees by the theory's empirical confirmation (Psillos [1999], Chapters 5, 6). In the context of GTR, one should recognize that the theory's success in weak- and medium-strength gravitational fields provides little support for the theory's accuracy

in strong gravitational fields such as near the Big Bang. But there is no reason to restrict the competitors of GTR to theories that someone on Earth has already proposed. The relevant set of competitors for GTR includes the set of theories that agree with GTR on all experiments to date, whether already entertained on Earth or not. This set might be infinite, might well be large, likely contains several members, and almost certainly has at least one member, a quantum theory of gravity. The set most likely has at least one member that resolves the singularities of GTR. Thus it is not at all clear why one should take GTR seriously in the strong-field regime near the supposed Big Bang singularity. Certainly, Einstein did not (Earman and Eisenstaedt [1999]).

11 Vicious God-of-the-gaps character

Regarding the argument from the Big Bang singularity to theism, Chris Isham notes 'its obvious susceptibility to the "God of the gaps" syndrome, in which God is relegated to filling in the blanks of an otherwise complete scientific theory' (Isham [1997], p. 378). Likewise John Earman warns that '[t]hose who want to find God in the big bang should beware of falling into the trap of relegating God to the diminishing interstices left by modern science' (Earman [1995], p. 209). Typical worries about 'God of the gaps' apologetic arguments involve the claim that there is a long history of appeals to special divine action to explain certain phenomena, but later natural explanations for such phenomena appeared, making the appeal to special divine action unnecessary and even foolish. Making an induction over this history, one is supposed to learn the lesson not to appeal to special divine action in new cases, lest one make religion look foolish yet again when the gaps close (Saunders [2002]).

On the other hand, it has been argued recently that worries about 'God of the gaps' arguments are overstated philosophically (Ratzsch [2001]; Snoke [2001]; Larmer [2002]). Del Ratzsch notes, for example, that the argument form is valid. These worries might also be overstated historically. Are they part of the same complex of distortion as the Huxley–Draper–White thesis that the characteristic mode of interaction between science and religion has been warfare (White [1896])? This claim has been refuted by modern historians of science (Lindberg and Numbers [1986]; Brooke [1991]; Olson [2004]). The warfare thesis, a large-scale generalization, keeps company with some specific claims that are simply false, such as that the medievals believed in a flat Earth (refuted in Russell [1991]; Grant [1994]). Given how many flaws have been diagnosed in the Huxley–Draper–White story by recent historians, one might wonder whether the definitive history of God-of-the-gaps arguments also has yet to be written.

Even if one concedes that some gaps arguments for theism might not be bad arguments, it remains clear that the Big Bang singularity argument *is* a bad argument from gaps to God. That is clear from the Destroyer *reductio*: there is no nonconventional relevant difference between the Big Bang singularity and the stellar gravitational collapse, and the latter surely has no theological significance and will likely disappear due to improved physics. Here I have assumed an intolerant attitude toward singularities. If one does tolerate singularities, then every moment of the

Big Bang model ($t > 0$) is preceded by earlier moments and again there is no call for extramundane explanation.

12 Fluctuating or inaccessible warrant

A related problem with arguing from the Big Bang to creation in time and hence theism is that such arguments depend crucially on various highly technical premises which most people cannot even entertain, much less evaluate. It follows that the vast majority of people, even educated ones, simply are not entitled to beliefs on the matter, apart from relying on the testimony of experts. But most people, even most educated people, cannot even reliably identify relevant experts. Most astronomers and physicists are not relevant experts, though they might well write popular books and make statements to the media on such issues. Supposing that one manages to identify relevant experts, the problem remains that their expert opinions will or should vary rather rapidly with the winds and waves of research fortune.

But should one's theology be affected much by the validity of Weak or Averaged Null Energy Condition assumed for singularity theorems? Many once-credible energy conditions have fallen by the wayside or are seriously threatened (Barcelo and Visser [2002]). It is now known that quantum field theory violates the local classical energy conditions, such as the Weak Energy Condition, though apparently quantum field theory still satisfies certain averaged energy conditions: energy density can be negative here and there, but not for very long and only with greater compensation of positive energy nearby (Ford and Roman [2001]; Barcelo and Visser [2002]; Ford [2003]). Nonminimally coupled classical scalar fields violate energy conditions, as does massive gravity (Visser [1998]). Are journalists, sociologists, homemakers, and truck drivers supposed to accept an argument whose premises are so technical that they cannot understand them, and so unstable that they could prove false in the next issue of *Physical Review D*? It is not clear why. Should the strength of one's belief in God depend on which factor ordering for the Hamiltonian constraint is correct in quantum gravity? Martin Bojowald recently wrote (Bojowald [2002]):

> Because of genuinely quantum geometrical effects the classical singularity is absent in those models in the sense that the evolution does not break down there, contrary to the classical situation where space time is inextendible. This effect is generic and does not depend on matter violating energy conditions, but it does depend on the factor ordering of the Hamiltonian constraint.

Most people have no idea what that means, and thus no idea what sort of plausibility to assign a particular factor ordering of the Hamiltonian constraint. If some of them believed in God because of the Big Bang singularity argument, must they now be able to refute Bojowald's choice of factor ordering in order to maintain that belief rationally? Even if the singularity argument could not be decisively refuted, could it

establish theism to a significant degree for anyone besides the few dozens of people expert in factor ordering in quantum gravity? It is unclear how. Could even the experts' judgments rationally remain stable enough to serve religious faith well? Probably not. But perhaps the choice of factor ordering is not so important after all (Date and Hossain [2005]); the fluctuations based on detailed technical premises continue.[44] If the singularity argument for theism does provide any warrant for theism, that warrant fluctuates wildly for experts and cannot be assessed by the laity.

Perhaps neither the singularity argument nor any other argument is the basis for religious faith. Craig holds to something like Reformed epistemology (private correspondence and Cowan *et al.* [2000]). Thus the religious believer does not need to read and understand every relevant paper in *Physical Review D* to maintain theistic belief rationally. Maybe this is the correct way to understand the warrant for theistic belief. But if the point of making the singularity argument is to provide an *argument* that rationally ought to persuade some people of theism, then Reformed epistemology is simply irrelevant to the task at hand. The fluctuating (for the experts) or inaccessible (for nonexperts) character of such warrant (if any) as an argument from the singularity to theism can provide, is another reason that theistic apologists ought to abandon this strategy.

13 Big Bang cosmology not especially congenial to faith

If the Big Bang singularity does not provide a good theistic argument, one might still consider whether Big Bang cosmology is somehow especially congenial to the eye of faith. Many hold that theism benefited from the victory of Big Bang cosmology over steady state cosmology. According to Francis Collins, '[t]he consequences of Big Bang theory for theology are profound. For faith traditions that describe the universe as having been created by God from nothingness (ex nihilo), this is an electrifying outcome' (Collins [2007], p. 66). As a sociological–historical claim, doubtless theism did so benefit, but is there any real philosophical advantage? Earman, noting that one can posit ideal boundary points 'at' singularities such as the Big Bang, reminds us that such boundary points are merely

> ideal elements. . . . Nothing prevents the theist from seeing God as operating at these ideal points. But since ideal points are not points of spacetime, the sense in which God can be said to cause or bring about the universe by operating at these points is very remote from the usual causal notions of science and everyday life that are concerned with connections between events in space and time. This is not to say that theistic talk about God creating the universe is illegitimate. But it is to say that such talk finds no special purchase in the big bang. *Even in models with no big bang and with time extending infinitely far into the past, ideal points corresponding to* $t = -\infty$ *could be attached to the spacetime manifold and God's helping hand could be seen at work there* (Earman [1995], p. 209; emphasis in the original).

Neither Big Bang cosmology nor some eternal-world cosmology has a beginning in the relevant sense of a first moment. Those who are willing to insist on a real first moment of (space)time, perhaps in accord with the Fourth Lateran Council, have the option of chopping off the front edge of the space-time model somewhere, whether in Big Bang cosmology or in some eternal-world cosmology, in an exercise of selective scientific antirealism. (Such a front edge presumably would be a Cauchy surface, everywhere space-like, but perhaps nearly null in places if one wishes. The question of just where to chop is vexing, however.) Along with C. J. S. Clarke (Clarke [1993], pp. 8, 9), Earman implicitly critiques such a proposal:

> Metaphysical considerations suggest that to be a serious candidate for describing actuality, a spacetime should be maximal. For example, for the Creative Force to actualize a proper subpart of a larger spacetime would seem to be a violation of Leibniz's principles of sufficient reason and plenitude. If one adopts the image of spacetime as being generated or built up as time passes then the dynamical version of the principle of sufficient reason would ask why the Creative Force would stop building if it is possible to continue (Earman [1995], p. 32).

On the other hand, Leibniz himself held (anachronism aside) that space-time is maximal toward the future, but it is not maximal toward the past (Alexander [1956], p. 76). God could have created the world such that events qualitatively identical to those of the first moment of the actual world would have been preceded by earlier events. 'But whether such an augmentation be reasonable and agreeable to God's wisdom, is another question, to which we answer in the negative; otherwise God would have made such an augmentation' (Alexander [1956], p. 76). Leibniz continues by quoting Horace about joining to a human head a horse's neck, a charming analogy, but not very helpful philosophically. Perhaps efficiency plays a role here in justifying nonmaximality toward the past, or perhaps the causal asymmetry or flow of time is relevant. In any case, whether one tacks on ideal beginning points, chops off the front edge of space-time to have a true first moment (and hence models space-time as a manifold with boundary), or takes a textbook space-time manifold without additions or subtractions, the 'initial' singularity and finite age of Big Bang cosmology provide no advantage over eternal universe cosmologies for theists, even when an apologetic agenda is dropped.

Given the many difficulties involved in arguing from Big Bang cosmology to creation in time, it is encouraging for theists and nontheists alike that Pope John Paul II has suggested caution in apologetic use of Big Bang cosmology:

> . . . some theologians, at least, should be sufficiently well-versed in the sciences to make authentic and creative use of the resources that the best-established theories may offer them. Such an expertise would prevent them from making uncritical and overhasty use for apologetic purposes of such recent theories as that of the 'Big Bang' in cosmology (John Paul II [1997], pp. M11, M12).

It should be noted that the concerns raised in this paper have little direct bearing on the persuasiveness of a teleological argument from cosmic fine tuning (Manson [2003]).

Acknowledgments

I thank William Lane Craig, Ernan McMullin, Don Howard, Samuel Conner, Scott Hawley, Jeff Koperski, Alex Skiles, John Keller, Jennifer Jensen, Dilkushi de Alwis Pitts, William Pitts, and Jan Pitts for discussions or other assistance, and perhaps others whom I cannot recall or name. All views and remaining errors are my own. An earlier version of this paper was presented at the Society of Christian Philosophers meeting at the University of Notre Dame in 2006.

Notes

1. The Lagrangian density of the Bach–Weyl theory is the square of the Weyl curvature tensor $C_{\rho\sigma\nu}{}^{\mu}$. The Weyl tensor in its natural habitat is a (1, 3) tensor and is a concomitant of the conformal metric density without the determinant $g =_{def} det(g_{\alpha\beta})$, so it is itself conformally invariant (Anderson [1967]; Wald [1984]). (Anderson has a misprint on p. 84 on this very point; one index of the Weyl tensor should be raised, as in $C_{\rho\sigma\nu}{}^{\mu}$, or else the index should be lowered with $\hat{g}_{\mu\nu}$ (not $g_{\mu\nu}$), thereby giving a densitized Weyl tensor.) Thus the Lagrangian density is (up to a constant factor) $\breve{g}^{\rho\alpha}\breve{g}^{\sigma\beta}C_{\rho\sigma\nu}{}^{\mu}C_{\alpha\beta\mu}{}^{\nu}$. The absence of g has been made manifest by the use of the inverse conformal metric tensor density $\hat{g}_{\mu\nu}$, which has density weight $\frac{1}{2}$, and the primordial (1, 3) form of the Weyl tensor. Much like the Maxwell electromagnetic kinetic term $\hat{g}^{\mu\alpha}\hat{g}^{\nu\beta}F^{\mu\nu}F^{\alpha\beta}$, the Bach–Weyl Lagrangian density is a weight 1 scalar density (as is required for a coordinate-invariant stationary action principle), but it does not depend on g, and so is manifestly conformally invariant.
2. For other space-time metrics, it can happen that a singularity manifests itself in the conformal metric density $\hat{g}_{\mu\nu}$ (Tod and Lübbe [unpublished]). The usual Big Bang is just not such a case. Thus the Big Bang singularity forms no insurmountable barrier in the Bach–Weyl theory, for example, because g is absent from that theory.
3. One might have other reasons for doubting that singularities exist as part of space-time if one emphasizes the analogy between gravity and other forces (Pitts and Schieve [2003]).
4. The point having been made, perhaps a brief explanation of the factor-ordering problem is appropriate (Komar [1979]). In classical physics, the Hamiltonian description involves various products of coordinates q and momenta p. In some theories, every term is just a power of q or of p, but not both. But what if there is a term involving powers of both q and p, as is true in Einstein's theory of gravity? Then the order in which they are written, such as p^2q^2 or q^2p^2 or $qppq$ for example, though of no importance classically, is of some importance in quantum theory.

References

Adler, R. J., Bjorken, J. D. and Overduin, J. M. [unpublished]: 'Finite Cosmology and a CMB Cold Spot', <arxiv.org/abs/gr-qc/0602102>.

Agnese, A. G. and Wataghin, A. [1971]: 'Scale Factor in Cosmological Homogeneous Models', *Lettere al Nuovo Cimento*, **1**, pp. 857–60.

Alexander, H. G., *ed.* [1956]: *The Leibniz–Clarke Correspondence*, Manchester: Manchester University Press.
Anderson, J. L. [1967]: *Principles of Relativity Physics*, New York: Academic Press.
Ashtekar, A. [unpublished]: 'Quantum Geometry and Gravity: Recent Advances', <arxiv.org/abs/gr-qc/0112038>. Report of the plenary talk at the 16th International Conference on General Relativity and Gravitation, held at Durban, S. Africa in July 2001.
Ashtekar, A., Pawlowski, T. and Singh, P. [2006a]: 'Quantum Nature of the Big Bang: An Analytical and Numerical Investigation', *Physical Review D*, **73**, pp. 1–33.
Ashtekar, A., Pawlowski, T. and Singh, P. [2006b]: 'Quantum Nature of the Big Bang: Improved Dynamics', *Physical Review D*, **74**, pp. 1–23.
Babak, S. V. and Grishchuk, L. P. [2003]: 'Finite-Range Gravity and Its Role in Gravitational Waves, Black Holes and Cosmology', *International Journal of Modern Physics D*, **12**, pp. 1905–59.
Barcelo, C. and Visser, M. [2002]: 'Twilight for the Energy Conditions?', *International Journal for Modern Physics D*, **11**, pp. 1553–60.
Bojowald, M. [2001a]: 'Absence of Singularity in Loop Quantum Cosmology', *Physical Review Letters*, **86**, pp. 5227–30.
Bojowald, M. [2001b]: 'The Inverse Scale Factor in Isotropic Quantum Geometry', *Physical Review D*, **64**, pp. 1–8.
Bojowald, M. [2002]: 'Isotropic Loop Quantum Cosmology', *Classical and Quantum Gravity*, **19**, pp. 2717–41.
Bojowald, M. and Hinterleitner, F. [2002]: 'Isotropic Loop Quantum Cosmology with Matter', *Physical Review D*, **66**, pp. 1–9.
Bojowald, M. and Kagan, M. [2006]: 'Singularities in Isotropic Non-Minimal Scalar Field Models', *Classical and Quantum Gravity*, **23**, pp. 4983–90.
Bondi, H. [1947]: 'Spherically Symmetrical Models in General Relativity', *Monthly Notices of the Royal Astronomical Society*, **107**, pp. 410–25.
Boulanger, N. and Esole, M. [2002]: 'A Note on the Uniqueness of D=4, N=1 Supergravity', *Classical and Quantum Gravity*, **19**, pp. 2107–24.
Boulware, D. G. and Deser, S. [1972]: 'Can Gravitation Have a Finite Range?', *Physical Review D*, **6**, pp. 3368–82.
Brooke, J. H. [1991]: *Science and Religion: Some Historical Perspectives*, Cambridge: Cambridge University Press.
Brown, H. R. [2005]: *Physical Relativity: Space-time Structure from a Dynamical Perspective*, New York: Oxford University Press.
Buchert, T. [2008]: 'Dark Energy from Structure: A Status Report', *General Relativity and Gravitation*, **40**, pp. 467–527.
Buchert, T. and Carfora, M. [2002]: 'Regional Averaging and Scaling in Relativistic Cosmology', *Classical and Quantum Gravity*, **19**, pp. 6109–45.
Callan, C., Dicke, R. H. and Peebles, P. J. E. [1965]: 'Cosmology and Newtonian Mechanics', *American Journal of Physics*, **33**, pp. 105–8.
Cartin, D. and Khanna, G. [2005]: 'Absence of Pre-Classical Solutions in Bianchi I Loop Quantum Cosmology', *Physical Review Letters*, **94**, pp. 1–4.
Catena, R., Pietroni, M. and Scarabello, L. [2007]: 'Einstein and Jordan Frames Reconciled: A Frame-Invariant Approach to Scalar–Tensor Cosmology', *Physical Review D*, **76**, pp. 1–11.

Cattoen, C. and Visser, M. [unpublished]: 'Cosmography: Extracting the Hubble Series from the Supernova Data', <arxiv.org/abs/gr-qc/0703122>.
Clarke, C. J. S. [1993]: *The Analysis of Space-Time Singularities*, Cambridge: Cambridge University Press.
Coley, A. A. and Pelavas, N. [2006]: 'Averaging Spherically Symmetric Spacetimes in General Relativity', *Physical Review D*, **74**, pp. 1–4.
Collins, F. S. [2007]: *The Language of God: A Scientist Presents Evidence for Belief*, New York: Free Press.
Copan, P. and Craig, W. L. [2004]: *Creation out of Nothing: A Biblical, Philosophical and Scientific Exploration*, Grand Rapids, MI: Baker Academic.
Cowan, S. B., Craig, W. L., Habermas, G. R., Feinberg, P. D., Frame, J. M. and Clark, K. J. [2000]: *Five Views on Apologetics*, Grand Rapids, MI: Zondervan.
Craig, W. L. [1979]: *The Kalām Cosmological Argument*, New York: Barnes and Noble.
Craig, W. L. and Smith, Q. [1993]: *Theism, Atheism, and Big Bang Cosmology*, Oxford: Clarendon Press.
Date, G. and Hossain, G. M. [2005]: 'Genericness of a Big Bounce in Isotropic Loop Quantum Cosmology', *Physical Review Letters*, **94**, pp. 1–4.
Deser, S. [1970]: 'Self-Interaction and Gauge Invariance', *General Relativity and Gravitation*, **1**, pp. 9–18.
Drees, W. B. [1990]: *Beyond the Big Bang: Quantum Cosmologies and God*, La Salle, IL: Open Court.
Dzhunushaliev, V. and Schmidt, H.-J. [2000]: 'New Vacuum Solutions of Conformal Weyl Gravity', *Journal of Mathematical Physics*, **41**, pp. 3007–15.
Earman, J. [1995]: *Bangs, Crunches, Whimpers, and Shrieks: Singularities and Acausalities in Relativistic Spacetimes*, New York: Oxford University Press.
Earman, J. [1996]: 'Tolerance for Spacetime Singularities', *Foundations of Physics*, **26**, pp. 623–40.
Earman, J. and Eisenstaedt, J. [1999]: 'Einstein and Singularities', *Studies in History and Philosophy of Modern Physics*, **30**, pp. 185–235.
Earman, J. and Mosterin, J. [1999]: 'A Critical Look at Inflationary Cosmology', *Philosophy of Science*, **66**, pp. 1–49.
Ellis, G. F. R. [1984]: 'Relativistic Cosmology: Its Nature, Aims and Problems', in B. Bertotti, F. de Felice and A. Pascolini (*eds)*, *General Relativity and Gravitation: Invited Papers and Discussion Reports of the 10th International Conference on General Relativity and Gravitation, Padua, July 3–8, 1983*, Dordrecht: D. Reidel, pp. 215–88.
Ellis, G. [2008]: 'Patchy Solutions', *Nature*, **452**, pp. 158–61.
Fang, J. and Fronsdal, C. [1979]: 'Deformations of Gauge Groups. Gravitation', *Journal of Mathematical Physics*, **20**, pp. 2264–2471.
Faraoni, V. and Nadeau, S. [2007]: '(Pseudo)issue of the Conformal Frame Revisited', *Physical Review D*, **75**, pp. 1–13.
Feynman, R. P., Morinigo, F. B., Wagner, W. G., Hatfield, B., Preskill, J. and Thorne, K. S. [1995]: *Feynman Lectures on Gravitation*, Reading, MA: Addison-Wesley. Original by California Institute of Technology, 1963.
Fiedler, B. and Schimming, R. [1980]: 'Exact Solutions of the Bach Field Equations of General Relativity', *Reports on Mathematical Physics*, **17**, pp. 15–36.
Fierz, M. and Pauli, W. [1939]: 'On Relativistic Wave Equations for Particles of Arbitrary Spin in an Electromagnetic Field', *Proceedings of the Royal Society (London) A*, **173**, pp. 211–32.

Flew, A. and Varghese, R. A. [2007]: *There Is a God: How the World's Most Notorious Atheist Changed His Mind*, New York: HarperOne.
Ford, L. H. [2003]: 'The Classical Singularity Theorems and Their Quantum Loopholes', *International Journal of Theoretical Physics*, **42**, pp. 1219–27.
Ford, L. H. and Roman, T. A. [2001]: 'Classical Scalar Fields and the Generalized Second Law', *Physical Review D*, **64**, pp. 1–12.
Gasperini, M. and Veneziano, G. [2003]: 'The Pre-Big Bang Scenario in String Cosmology', *Physics Reports*, **373**, pp. 1–212.
Gilkey, L. [1959]: *Maker of Heaven and Earth: The Christian Doctrine of Creation in the Light of Modern Knowledge*, Garden City, NY: Doubleday and Company. Page reference is to the 1965 Anchor Books reprint.
Giulini, D. [2007]: 'Some Remarks on the Notions of General Covariance and Background Independence', in E. Seiler and I. O. Stamatescu (*eds), Approaches to Fundamental Physics: An Assessment of Current Theoretical Ideas*, Lecture Notes in Physics, vol. 721, New York: Springer, pp. 105–22.
Glymour, C. [1973]: 'Topology, Cosmology, and Convention', in P. Suppes (*ed.), Space, Time, and Geometry*, Dordrecht: D. Reidel, pp. 193–216.
Grant, E. [1994]: *Planets, Stars and Orbs: The Medieval Cosmos, 1200–1687*, Cambridge: Cambridge University Press.
Grant, E. [1996]: *The Foundations of Modern Science in the Middle Ages: Their Religious, Institutional, and Intellectual Contexts*, Cambridge: Cambridge University Press.
Grünbaum, A. [1989]: 'The Pseudo-Problem of Creation in Physical Cosmology', *Philosophy of Science*, **56**, pp. 373–94.
Gupta, S. N. [1954]: 'Gravitation and Electromagnetism', *Physical Review*, **96**, pp. 1683–5.
Helm, P. [1997]: 'Eternal Creation: The Doctrine of the Two Standpoints', in C. E. Gunton (*ed.*), *The Doctrine of Creation: Essays in Dogmatics, History and Philosophy,* Edinburgh: T&T Clark, pp. 29–46.
Henneaux, M. and Teitelboim, C. [1992]: *Quantization of Gauge Systems*, Princeton, NJ: Princeton University Press.
Hyman, A. and Walsh, J. J. [1973]: *Philosophy in the Middle Ages: The Christian, Islamic, and Jewish Traditions*, 2nd edition, Indianapolis, MN: Hackett.
Infeld, L. and Schild, A. [1945]: 'A New Approach to Kinematic Cosmology', *Physical Review*, **68**, pp. 250–72.
Isham, C. J. [1997]: 'Creation of the Universe as a Quantum Process', in R. J. Russell, W. R. Stoeger and G. V. Coyne (eds), *Physics, Philosophy, and Theology: A Common Quest for Understanding*, 3rd edition, Vatican City State: Vatican Observatory/Notre Dame, IN: University of Notre Dame, pp. 375–408.
Jackson, J. D. [1975]: *Classical Electrodynamics*, 2nd edition, New York: Wiley.
Jaki, S. L. [1980]: *Cosmos and Creator.* Edinburgh: Scottish Academic Press.
John Paul II [1997]: 'To the Reverend George V. Coyne, S. J., Director of the Vatican Observatory, 1 June 1988', in R. J. Russell, W R. Stoeger and G. V. Coyne (eds), *Physics, Philosophy, and Theology: A Common Quest for Understanding*, 3rd edition, Vatican City State: Vatican Observatory/Notre Dame, IN: University of Notre Dame, pp. **15**, 16, M1–14.
Kaloper, N. and Olive, K. A. [1998]: 'Singularities in Scalar–Tensor Cosmologies', *Physical Review D*, **57**, pp. 811–22.

Kazanas, D. and Mannheim, P. D. [1991]: 'General Structure of the Gravitational Equations of Motion in Conformal Weyl Gravity', *Astrophysical Journal Supplement Series*, **76**(2), pp. 431–53.
Klein, O. [1971]: 'Arguments Concerning Relativity and Cosmology', *Science*, **171**, pp. 339–45.
Komar, A. [1979]: 'Consistent Factor Ordering of General-Relativistic Constraints', *Physical Review D*, **20**, pp. 830–3.
Kragh, H. [2004]: *Matter and Spirit in the Universe: Scientific and Religious Preludes to Modern Cosmology*, London: Imperial College.
Kraichnan, R. H. [1955]: 'Special-Relativistic Derivation of Generally Covariant Gravitation Theory', *Physical Review*, **98**, pp. 1118–22.
Kuhn, T. S. [1978]: *Black-Body Theory and the Quantum Discontinuity, 1894–1912*, Oxford: Clarendon Press.
Kukla, A. [1998]: *Studies in Scientific Realism*, New York: Oxford University Press.
Küng, H. [2007]: *The Beginning of All Things: Science and Religion.* Grand Rapids, MI: William B. Eerdmans. Translated by John Bowden.
Larmer, R. [2002]: 'Is There Anything Wrong with "God of the Gaps" Reasoning?', *International Journal for Philosophy of Religion*, **52**, pp. 129–42.
Laudan, L., Donovan, A., Laudan, R., Barker, P., Brown, H., Leplin, J., Thagard, P. and Wykstra, S. [1986]: 'Scientific Change: Philosophical Models and Historical Research', *Synthese*, **69**, pp. 141–223.
Layzer, D. [1954]: 'On the Significance of Newtonian Cosmology', *The Astronomical Journal*, **59**, pp. 268–70.
Leith, J. H., *ed.* [1982]: *Creeds of the Churches: A Reader in Christian Doctrine from the Bible to the Present*, 3rd edition, Louisville, KY: John Knox.
Levy-Leblond, J.-M. [1990]: 'Did the Big Bang Begin?', *American Journal of Physics*, **58**, pp. 156–9.
Lindberg, D. C. and Numbers, R. L. *eds* [1986]: *God & Nature: Historical Essays on the Encounter between Science and Religion*, Berkeley, CA: University of California.
Magnano, G. and Sokołowski, L. M. [1994]: 'Physical Equivalence between Nonlinear Gravity Theories and a General-Relativistic Self-Gravitating Scalar Field', *Physical Review D*, **50**, pp. 5039–59.
Malament, D. [1977]: 'Observationally Indistinguishable Space-Times', in J. Earman C. Glymour and J. Stachel (*eds*), *Foundations of Space-Time Theories*, Minneapolis, MN: University of Minnesota, pp. 61–80.
Manson, N. A., *ed.* [2003]: *God and Design: The Teleological Argument and Modern Science*, London: Routledge.
Mars, M. and Zalaletdinov, R. M. [1997]: 'Space-Time Averages in Macroscopic Gravity and Volume-Preserving Coordinates', *Journal of Mathematical Physics*, **38**, pp. 4741–57.
Mavrodes, G. I. [1989]: 'Enthusiasm', *International Journal for Philosophy of Religion*, **25**, pp. 171–86.
McCabe, G. [2004]: 'The Structure and Interpretation of Cosmology: Part I – General Relativistic Cosmology', *Studies in History and Philosophy of Modern Physics*, **35**, pp. 549–95.
McCrea, W H. [1955]: 'On the Significance of Newtonian Cosmology', *The Astronomical Journal*, **60**, pp. 271–4.

McMullin, E. [1981a]: 'How Should Cosmology Relate to Theology?', in A. R. Peacocke (*ed.)*, *The Sciences and Theology in the Twentieth Century*, Notre Dame, IN: University of Notre Dame, pp. 17–57.
McMullin, E. [1981b]: 'Is Philosophy Relevant to Cosmology?', *American Philosophical Quarterly*, **18**, pp. 177–89.
Misner, C. W [1969]: 'Absolute Zero of Time', *Physical Review A*, **186**, pp. 1328–33.
Misner, C., Thorne, K. and Wheeler, J. A. [1973]: *Gravitation*, New York: Freeman.
Morriston, W [2003]: 'Must Metaphysical Time Have a Beginning?', *Faith and Philosophy*, **20**, pp. 288–306.
Narlikar, J. V. [1992]: 'The Concepts of "Beginning" and "Creation" in Cosmology', *Philosophy of Science*, **59**, pp. 361–71.
Ogievetsky, V. I. and Polubarinov, I. V. [1965]: 'Interacting Field of Spin 2 and the Einstein Equations', *Annals of Physics*, **35**, pp. 167–208.
Olson, R. G. [2004]: *Science & Religion, 1450–1900: From Copernicus to Darwin*, Baltimore, MD: Johns Hopkins University Press.
Papapetrou, A. [1948]: 'Einstein's Theory of Gravitation and Flat Space', *Proceedings of the Royal Irish Academy A*, **52**, pp. 11–23.
Pitts, J. B. [2006]: 'Absolute Objects and Counterexamples: Jones–Geroch Dust, Torretti Constant Curvature, Tetrad-Spinor, and Scalar Density', *Studies in History and Philosophy of Modern Physics*, **37**, pp. 347–71.
Pitts, J. B. and Schieve, W. C. [2001]: 'Slightly Bimetric Gravitation', *General Relativity and Gravitation*, **33**, pp. 1319–50.
Pitts, J. B. and Schieve, W. C. [2003]: 'Nonsingularity of Flat Robertson–Walker Models in the Special Relativistic Approach to Einstein's Equations', *Foundations of Physics*, **33**, pp. 1315–21.
Pitts, J. B. and Schieve, W. C. [2007]: 'Universally Coupled Massive Gravity', *Theoretical and Mathematical Physics*, **151**, pp. 700–17.
Pius III [1952]: 'Modern Science and the Existence of God', *The Catholic Mind*, **49,** pp. 182–92. Address to the Pontifical Academy of Science, November **22**, 1951.
Plantinga, A. [2000]: *Warranted Christian Belief*, New York: Oxford University Press.
Polkinghorne, J. [1994]: *Science and Christian Belief*, London: SPCK.
Psillos, S. [1999]: *Scientific Realism: How Science Tracks Truth*, London: Routledge.
Pullin, J. [2003]: 'Canonical Quantization of General Relativity: The Last 18 Years in a Nutshell', in M. Novello (*ed.), Proceedings of the Xth Brazilian School on Cosmology and Gravitation: AIP Conference Proceedings* 668, American Institute of Physics, pp. 141–53.
Quinn, P. L. [1993]: 'Creation, Conservation, and the Big Bang', in J. Earman A. I. Janis, G. J. Massey and N. Rescher (*eds)*, *Philosophical Problems of the Internal and External Worlds: Essays on the Philosophy of Adolf Grünbaum*, Pittsburgh: University of Pittsburgh/Konstanz: Universitätsverlag Konstanz, pp. 589–612.
Ratzsch, D. [2001]: *Nature, Design, and Science: The Status of Design in Natural Science*, Albany, NY: State University of New York.
Redhead, M. [1996]: Review of *Theism, Atheism, and Big Bang Cosmology* by William Lane Craig and Quentin Smith, *British Journal for the Philosophy of Science*, **47**, pp. 133–6.
Rosen, N. [1940]: 'General Relativity and Flat Space. I, II', *Physical Review*, **57**, pp. 147–50, 150–3.
Ross, H. [1991]: *The Fingerprint of God*, 2nd edition, Orange, CA: Promise.

Ross, H. [2001]: *The Creator and the Cosmos: How the Greatest Scientific Discoveries of the Century Reveal God*, 3rd edition, Colorado Springs: NavPress.

Russell, J. B. [1991]: *Inventing the Flat Earth: Columbus and Modern Historians*, Westport, CN: Praeger.

Russell, R. J. [1996a]: 'Finite Creation Without a Beginning', in R. J. Russell, N. Murphey and C. J. Isham (*eds)*, *Quantum Cosmology and the Laws of Nature: Scientific Perspectives on Divine Action*, 2nd edition, Vatican City State: Vatican Observatory/Berkeley, CA: Center for Theology and the Natural Sciences/Notre Dame, IN: University of Notre Dame, pp. 291–325.

Russell, R. J. [1996b]: '*t* = 0: Is It Theologically Significant?', in W. M. Richardson and W. J. Wildman (*eds.)*, *Religion & Science: History, Method, Dialogue*, New York: Routledge, pp. 201–24.

Santiago, D. I. and Silbergleit, A. S. [2000]: 'On the Energy–Momentum Tensor of the Scalar Field in Scalar–Tensor Theories of Gravity', *General Relativity and Gravitation*, **32**, pp. 565–81.

Saunders, N. [2002]: *Divine Action & Modern Science*, Cambridge: Cambridge University Press.

Schimming, R. and Schmidt, H.-J. [1990]: 'On the History of Fourth Order Metric Theories of Gravitation', *NTM Schriftenreihe für Geschichte der Naturwissenschaften, Technik und Medizin*, **27**, pp. 41–8.

Schücking, E. L. [1967]: 'Cosmology', in J. Ehlers (*ed.)*, *Relativity Theory and Astrophysics: 1. Relativity and Cosmology*, Lectures in Applied Mathematics, vol. 8, Providence, RI: American Mathematical Society, pp. 218–36.

Sexl, R. U. [1967]: 'Theories of Gravitation', *Fortschritte der Physik*, **15**, 269–307.

Sklar, L. [1985]: 'Do Unborn Hypotheses Have Rights?', in *Philosophy and Spacetime Physics*, Berkeley, CA: University of California.

Smith, Q. [1992]: 'A Big Bang Cosmological Argument for God's Nonexistence', *Faith and Philosophy*, **9**, pp. 217–37.

Smith, Q. [2000]: 'Problems with John Earman's Attempt to Reconcile Theism with General Relativity', *Erkenntnis*, **52**, pp. 1–27.

Smoller, J. and Temple, B. [2003]: 'Shock-Wave Cosmology Inside a Black Hole', *Proceedings of the National Academy of Sciences of the United States of America*, 100, pp. 11216–8.

Snoke, D. [2001]: 'In Favor of God-of-the-Gaps Reasoning', *Perspectives on Science & Christian Faith*, **53**, pp. 152–8.

Sober, E. [2003]: 'The Design Argument', in N. A. Manson (*ed.), God and Design: The Teleological Argument and Modern Science*, London: Routledge, pp. 27–54.

Sober, E. [2007]: 'What is Wrong with Intelligent Design?', *Quarterly Review of Biology*, **82**, pp. 3–8.

Sotiriou, T. P., Faraoni, V and Liberati, S. [2008]: 'Theory of Gravitation Theories: A No-progress Report', *International Journal of Modern Physics D*, **17**, pp. 399–423.

Stanford, P. K. [2001]: 'Refusing the Devil's Bargain: What Kind of Underdetermination Should We Take Seriously?', *Philosophy of Science*, **68**, pp. S1–12. PSA 2000 Proceedings.

Stanford, P. K. [2006]: *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*, New York: Oxford University Press.

Sundermeyer, K. [1982]: *Constrained Dynamics: With Applications to Yang–Mills Theory, General Relativity, Classical Spin, Dual String Model*, Lecture Notes in Physics, vol. 169, Berlin: Springer.

Tauber, G. E. [1967]: 'Expanding Universes in Conformally Flat Coordinates', *Journal of Mathematical Physics*, **8**, pp. 118–23.

Thijssen, J. M. M. H. [1998]: 'Eternity of the World, Medieval Views of', in E. Craig (*ed.*), *Routledge Encyclopedia of Philosophy*. London: Routledge. Retrieved November 13, 2004, from <www.rep.routledge.com/article/B039>.

Tod, P. and Lübbe, C. [unpublished]: 'An Extension Theorem for Conformal Gauge Singularities', <arxiv.org/abs/0710.5552>.

van Fraassen, B. [1989]: *Laws and Symmetry*, Oxford: Clarendon Press.

van Nieuwenhuizen, P. [1973]: 'On Ghost-Free Tensor Lagrangians and Linearized Gravitation', *Nuclear Physics B*, **60**, pp. 478–92.

Veltman, M. [1981]: 'Quantum Theory of Gravitation', in *Les Houches XXVIII 1975: Methods in Field Theory*, Amsterdam: North-Holland, pp. 265–327.

Visser, M. [1998]: 'Mass for the Graviton', *General Relativity and Gravitation*, **30**, pp. 1717–28.

Wald, R. M. [1984]: *General Relativity*, Chicago, IL: University of Chicago Press.

Weinberg, S. [1965]: 'Photons and Gravitons in Perturbation Theory: Derivation of Maxwell's and Einstein's Equations', *Physical Review*, 138, pp. B988–1002.

Weinberg, S. [1972]: *Gravitation and Cosmology*, New York: Wiley.

Weinstein, S. [1996]: 'Strange Couplings and Space-Time Structure', *Philosophy of Science*, **63**, pp. S63–70. Proceedings of the 1996 Biennial Meetings of the Philosophy of Science Association. Part I: Contributed Papers.

White, A. D. [1896]: *A History of the Warfare of Science with Theology in Christendom*, Buffalo: Prometheus, 1993 reprint.

Will, C. M. [1993]: *Theory and Experiment in Gravitational Physics*, revised edition, Cambridge: Cambridge University Press.

Zalaletdinov, R. M. [1992]: 'Averaging out the Einstein Equations', *General Relativity and Gravitation*, **24**, pp. 1015–31.

3

On Non-Singular Spacetimes and the Beginning of the Universe

William Lane Craig and James D. Sinclair

1.0 Introduction

The most widely discussed argument in natural theology today is the so-called *kalām* cosmological argument, which originated in attempts on the part of certain ancient philosophers to rebut Aristotle's doctrine of the past eternity of the universe.[1] The argument assumed major importance in mediaeval Islamic theology, from which its name derives. In his *Kitab al-Iqtisad* the medieval Muslim theologian al-Ghazali presented the following simple syllogism in support of the existence of a Creator: "Every being which begins has a cause for its beginning; now the world is a being which begins; therefore, it possesses a cause for its beginning" [al-Ghazali, 1962, pp. 15–16]. In defense of the second premiss, Ghazali offered various philosophical arguments to show the impossibility of an infinite regress of temporal phenomena and, hence, of an infinite past. The limit at which the finite past terminates Ghazali calls "the Eternal" [al-Ghazali, 1963, p. 32], which he evidently takes to be a state of timelessness. Given the truth of the first premiss, the finite past must, therefore, "stop at an eternal being from which the first temporal being should have originated" [al-Ghazali 1963, p. 33].

The argument, then, is disarmingly simple:

1. Everything that begins to exist has a cause.
2. The universe began to exist.
3. Therefore, the universe has a cause.

Conceptual analysis of what it means to be a cause of the universe then seeks explore the relevance of this conclusion for theism. What makes the *kalām* cosmological argument scientifically interesting is that contemporary proponents of the argument have claimed support from current astrophysical cosmogony in defense of (2). A number of recent critics of the argument, however, have denied that contemporary physical cosmology lends any significant support to (2) because the singular origin of spacetime predicted by the standard model is an artifact of General Relativity (hereafter GTR) which may well be resolved by quantum cosmological models.

For example, Bradley Monton states,

> The big bang hypothesis is true given the assumption that GTR is true, but we don't know that the big bang hypothesis is true of the actual universe. The big bang theory doesn't take into account quantum theory, and that gives us reason not to believe the big bang theory ….. Because the physics doesn't tell us what happens once we trace the history of the universe backwards in time to these high energies, we don't even know if there's a big bang at all [Monton, forthcoming].

Monton appeals specifically to the Steinhardt-Turok cyclic ekpyrotic model as an example of a cosmological model which "is compatible with the universe having been in existence forever" [Monton, forthcoming]. Similarly, Graham Oppy complains that we are "far from having good reason" to suppose that quantum gravitational replacements of the standard model will feature an absolute beginning of the physical universe [Oppy, 2006, p. 146]. Undoubtedly, the most extensively developed criticism of this sort comes J. Brian Pitts in his wide-ranging critique of what he styles "the singularity argument for theism" [Pitts, 2008, p. 677]. His critique therefore merits careful scrutiny.

It is a curious feature of Pitts' lengthy critique that he never actually provides an explicit statement of the target singularity argument for theism. At first blush, one might think that the *kalām* cosmological argument just is the singularity argument for theism. But a moment's reflection shows that this cannot be right, for at most, evidence for an initial cosmological singularity might be adduced in support of premise (2). It is immediately obvious, then, that "the singularity argument for theism" is a misnomer, for the singularity argument is not an argument *for theism*. Rather it is an argument for (2), which is a straightforward, physical statement, the likes of which is discussed in any text on astronomy and astrophysics and which does not posit any supernatural entity. What is really at issue, then, is some sort of singularity argument for the beginning of the universe.[2] The recognition of this fact is highly significant, since it implies that all of Pitts' worries about "God-of-the-gaps" reasoning and the impenetrability of natural theology [Pitts, 2008, pp. 696–700] are simply irrelevant to the success of the so-called singularity argument for the beginning of the universe and so may be left aside. In what follows, then, we shall have very little to say of theology. This is, we think, all for the better, for the question of whether contemporary cosmology furnishes good grounds for thinking that the universe began to exist is of more than sectarian interest.

2.0 The singularity argument formulated

Our concern, then, is with some supposed singularity argument for the beginning of the universe. In order to discern the alleged failings of that argument, we need a clear formulation of the argument. This Pitts unfortunately does not provide, leaving us to reconstruct it as best we can. The key to reconstructing the so-called singularity argument appears to be Pitts' summarizing statement:

> Whether one is tolerant or intolerant toward singularities, it turns out that there is no first moment (unless one is installed by hand), because every moment is preceded by earlier moments. Thus in the relevant sense for the Kalam argument to be valid and to make sense for a sufficiently broad collection of physical theories, there is no beginning implied by physics, and so premise 2 might be false, as far as physics can show. In order for the Big Bang singularity to provide a good theistic argument, the singularity must be well enough behaved to be a real and intelligible part of space-time and badly enough behaved that it cannot have a past. Satisfying both conditions seems difficult and unlikely to be achieved [Pitts, 2008, p. 689].

This passage might suggest the following reconstruction:

4. The universe began to exist iff there was a first moment of time.
5. If the initial cosmological singularity was a real and intelligible part of space-time, then there was a first moment of time.
6. The initial cosmological singularity was a real and intelligible part of space-time.
7. Therefore, the universe began to exist.

The problem with this suggested reconstruction of the argument, however, is that the argument would then be a straw man which has been defended by no one and would hardly require 21 pages of sustained argumentation to refute. Singularities of general relativistic spacetimes are not spacetime points but, at most, points, whether real or ideal, which can be attached to spacetime as boundary points. As such, they are not parts of spacetime, as (6) asserts. Indeed, on the customary approach to singularities, they are not even accorded that much reality. Rather a spacetime is said to be singular if it is essentially inextendible and so geodesically incomplete [Earman, 1995, pp. 28–37]. The question Pitts raises is whether an initial cosmological singularity, either in the sense of a past boundary point or of the incompleteness of past geodesic half-curves, gives grounds for thinking that the universe began to exist. Asking the question in this way renders intelligible Pitts' further worry whether some as yet unknown theory might not resolve the singularities of classical spacetime—a worry which would become superfluous with respect to the above proposed reconstruction—, for even if an initial cosmological singularity implies the beginning of the universe, the question remains whether spacetime really is singular.

Accordingly, we might offer the following reconstruction of the supposed singularity argument for the beginning of the universe:

8. If spacetime is singular, the universe began to exist.
9. Spacetime is singular.
10. Therefore, the universe began to exist.

Pitts may then be understood to deny (8) on the grounds that, whether the initial singularity is understood as a boundary point to the past or as the past geodesic incompleteness of spacetime, there is no first moment of the universe's existence, which he takes to be a necessary condition of the universe's beginning to exist, and

to undercut (9) by appealing to as yet undiscovered theories which remove the singularity. This seems to us a sympathetic reconstruction of the alleged singularity argument for the beginning of the universe which also makes good sense of Pitts' two-pronged response [Pitts, 2008, p. 682].

3.0 The singularity argument assessed

Even so, it must be said that the singularity argument for the beginning of the universe is still a straw man. For the evidence that the universe began to exist does not depend on evidence for a *singular* beginning of spacetime. We have elsewhere discussed a variety of non-singular cosmogonic models and have noted the difficulties in plausibly tracing out an infinite past [Copan and Craig, 2004; Craig, 2006; Craig and Sinclair, 2009]. We shall have more to say of this in the sequel, when examining what support contemporary physical cosmology lends to (2).

3.1 Assessment of premise (8)

Taking the argument on its own terms, however, let us first examine Pitts' objection to (8). Despite his flirtation with metric conventionalism, Pitts recognizes that our universe is characterized by an objective temporal metric. A Friedmann-Robertson-Walker universe, being bounded by an initial cosmological singularity, is metrically finite in the past. Following Smith [1985], we can say plausibly that time begins to exist if for any arbitrarily designated, non-zero, finite interval of time, there are only a finite number of isochronous intervals earlier than it; or, alternatively, time begins to exist if for some non-zero, finite temporal interval there is no isochronous interval earlier than it. That condition is satisfied in standard model spacetime. So why is the metrical finitude of the past not sufficient for time's beginning to exist? Pitts' answer is that "Within the Robertson-Walker cosmological space-time for $t > 0$ (which is to say, always), one can explain each moment in terms of an earlier one. Thus there is no beginning required and premise 2 might be false, as far as physics can tell" [Pitts, 2008, p. 688]. Pitts' point is not that each instantaneous slice of spacetime can be immanently explained in terms of a prior slice, thereby obviating the need for a transcendent cause of the universe's beginning to exist;[3] rather his point is that since $t = 0$ is at most a boundary point of spacetime, there is no first instant of time and thus no beginning of time, even if past time is finite, and, hence, no beginning of the universe.

Pitts' objection presupposes that *beginning to exist* entails *having a beginning point*. But why should we think that? Pitts offers three reasons why construing the universe's beginning to exist in topological rather than metrical terms is theologically advantageous for the biblical doctrine of creation [Pitts, 2008, pp. 680–1]. But such a concern is foreign, not only to natural theology, but to the religiously neutral question of whether the universe began to exist. What is relevant is philosophical considerations that have purchase with philosophers and cosmologists regardless of their religious persuasion. What is advantageous to the biblical doctrine of creation does not and

should not enter into the question of what it is to begin to exist.[4] Indeed, given the above intuitively plausible sufficient conditions for time's beginning to exist, Pitts' scenario of a metrically finite past lacking a first instant may be taken to provide good reason to conclude that *beginning to exist* does not entail *having a beginning point*.[5]

Moreover, Pitts' entailment would commit us to the reality of points. But whether space and time really are composed of an actual, nondenumerable infinity of points, rather than simply modeled as such in GTR, is surely a question to be settled by argument and evidence, not merely assumed, as Pitts does [Pitts, 2008, p. 680].

Furthermore, Pitts' own Cosmic Destroyer argument [Pitts, 2008, pp. 692–3] seems to undermine his presumed entailment. For if *beginning to exist* entails *having a beginning point,* then*, pari passu, ceasing to exist* entails *having an ending point.* Since a black hole singularity (or a terminal cosmological singularity) is at most a boundary point of spacetime rather than a part of spacetime, any object which falls into a black hole has no ending point of its existence. Nevertheless, it does cease to exist. Robert Geroch explains,

> Recall there is no event available to [an observer headed toward a singularity] on the singularity itself. In particular, there is no possibility for a further extension of [the observer's] world-line after it hits the singularity. What, then, does 'the world-line of [an observer] hits the singularity' mean physically? Mathematically, what happens is that this world-line just stops. Physically, this would mean that [the observer] is 'snuffed out of existence'; after some finite time according to himself, he ceases to exist in space-time [Geroch, 1978, p. 194].

This isn't simply rhetorical gloss on Geroch's part. The observer would indeed be 'snuffed out of existence' despite the lack of a genuine ending point to the spacetime. If, then, the universe ceases to exist in the Big Crunch, parity requires that it began to exist in the Big Bang, *Q.E.D.*[6]

Finally, it deserves to be pointed out that Pitts' insistence that the crucial consideration as to whether the universe began to exist is the necessity of a first instant lands him squarely in the ancient Greek paradoxes of stopping and starting [see Sorabji, 1983, pp. 403–21]. Ancient thinkers like Parmenides argued that if an object O is at rest at time t, it is impossible for O to begin to move, since for any time $t' > t$, if O is in motion at t', then there is a time $t < t^* < t'$ at which O is already in motion. Hence, nothing can ever begin to move. If we assume the continuity of time and space, the solution to Parmenides' puzzle is that O can begin to move without there being a beginning point of its motion. Pitts' demand for topological closure would force upon us the absurd conclusion that nothing ever begins to move. But if we allow that *beginning to move* does not entail *having a beginning point* of motion, then, generalizing, neither should we demand that in beginning to exist the universe must have a beginning point of its existence.

We need, then, some compelling reason to think that the admitted finitude of past time is not sufficient for time's having begun to exist, for thinking that a first instant of time is a necessary condition as well. Pitts' relevant remarks in this regard come in

response to Copan and Craig's definitions which they provide in preparation for their defense of the premiss

3*. The temporal series of past physical events was not beginningless.

They explain,

> By a 'physical event,' we mean any change occurring within the space-time universe. Since any change takes time, there are no instantaneous events. Neither could there be an infinitely slow event, since such an 'event' would in reality be a changeless state. Therefore, any event will have a finite, nonzero duration. In order that all the events be of equal duration, we arbitrarily stipulate some event as our standard.
>
> Taking as our point of departure the present standard event, we consider any series of such standard events ordered according to the relation *earlier than*. The question is whether this series of events had a beginning or not. By a 'beginning,' one means a first standard event. It is therefore not relevant whether the temporal series had a beginning point (a first temporal instant). The question is whether there was in the past an event occupying a nonzero, finite temporal interval that was absolutely first, not preceded by any equal interval [Copan and Craig, 2004, p. 199].

Pitts' objection to this *Ansatz* is surprising. He writes,

> The Bach-Weyl theory of gravity shows that a physical theory need not even define the length of a curve. In scalar-tensor theories or other theories with multiple metrics, age might be radically ambiguous. Thus, only a topological notion of 'beginning' in terms of a first moment is available if the Kalam argument is intended to yield a necessarily nonvacuous necessary truth …..
>
> Copan and Craig perhaps take their Kalam argument to express a necessary truth that applies nonvacuously to all possible physical theories. However, their criterion for a beginning is meaningless for a Bach-Weyl theory because 'equal duration' is meaningless in that theory ….. Given that neither existence nor uniqueness of a metric (for timelike curves) holds necessarily, the natural move is to adopt a topological rather than metrical notion of beginning. Thus a first moment *is* the point that needs to be addressed [Pitts, 2008, pp. 677, 682].

Pitts' presumption seems to be that Copan and Craig take, not only the premises of the *kalām* cosmological argument, but more particularly their proffered criterion for there being a beginning of the series of past events to be metaphysically necessary truths. Since there are possible worlds characterized by metric conventionalism—that is to say, worlds in which there is no objective fact of the matter whether or not non-nested temporal intervals are isochronous—, the proffered criterion for a beginning would be inapplicable to such worlds, for there just is no fact of the matter in such worlds whether any other temporal interval is equal to the duration of the arbitrarily

stipulated, standard event. Pitts recognizes, as mentioned, that the actual world is characterized by an objective metric; but he insists that "a doctrine of creation needs to be modally rich enough to accommodate the possibility of God's creating worlds with physical laws without a unique or preferred metrical structure to license an answer of 'finite' or 'infinite' age" [Pitts, 2008 p. 681].

Pitts' presumption is wrong. The proffered criterion was intended only to apply to the actual world. There is no intention or need to seek its application to other metaphysically possible words governed by different laws of nature.[7] Thus, the proffered criterion for a beginning of the series of past events is adequate to its purpose. It states a sufficient condition for time's having begun to exist.

In sum, Pitts' attempt to discredit an intuitively plausible sufficient condition of time and the universe's having begun to exist in terms of metrical past finitude rests upon an ungrounded presumption about the modal status of the proffered condition and would substitute a topological condition which, if accepted, would have highly implausible consequences. Hence, he presents no serious objection to (8).

3.2 Assessment of premise (9)

What, then, about (9)? If (9) is true, then, as we have seen above, we do have good reason to infer that (10) is true, that the universe began to exist, which is the second premise of the *kalām* cosmological argument. Unfortunately, Pitts seems to assume that the evidential support for premise (2) depends crucially on the cogency of his singularity argument, so that if the warrant for (9) is undercut, so is the warrant for (2). But this assumption is mistaken. Premise (2) does not presuppose or require the truth of (9) or the cogency of the singularity argument. Evidence of a non-singular beginning would be quite sufficient to warrant belief in (2).

Of course, if (9) is unwarranted, then we must ask what support physical cosmology lends to (2). Pitts is worried by the fact that

> As one sees all the time in papers on quantum gravity, most people who work on quantum gravity take for granted that the Big Bang singularity is an artifact of incomplete physical understanding and expect or hope that uniting gravity with quantum mechanics in some kind of quantum gravity will resolve the singularity into some well-defined situation that admits extrapolation to still earlier times, *ad infinitum* [Pitts, 2008, p. 688].[8]

The crucial consideration here will not be Pitts' worry that a successful theory of quantum gravity will resolve the initial cosmological singularity featured in the standard model into a situation which is physically well-defined, but rather whether, as he puts it, that situation "admits extrapolation to still earlier times, *ad infinitum*." Since it is that consideration, rather than the truth of (9), that is crucial to the evidence for (2), let us turn to an examination of that consideration. Do models featuring an era "before the Big Bang" imply a beginningless past?

4.0 Physical evidence for premise (2)

We shall now consider some contemporary cosmogonies.[9] The material in sections 4.2–6 is not intended as an inductive attempt to show that all contemporary cosmogonies may lend support to (2). Rather, it is intended to support the view evinced in [Craig & Sinclair 2009] that there are general principles regarding cosmogonies that yield a taxonomy of models based on the models' expansion behavior which speaks to the question of temporal origin. We will hold that there is a genuine unique objective past (temporally ordered), with evolutionary continuity to the present. This preserves our veridical experience of time[10] and permits theories of evolution (including Darwin) elsewhere in physics to be understood on a realist footing, as opposed to some type of illusion. Thus, for example, if a particular theory denies temporality as a 'fundamental' feature of the universe, but preserves the notion as a quasi-classical 'approximation', we will take the quasi-classical history as real and suggest the more general formalism should be interpreted instrumentally.

Further, as Pitts seems to assert without evidence that a pre-Big-Bang timeline would, of necessity, be past eternal, we shall content ourselves to providing a few counterexamples as opposed to attempting to prove that *all* pre-Big-Bang models fail to demonstrate past eternality.

4.1 Models, methodology, and their use

In the sections following this one, we shall consider the evidence in favor of the proposition that the universe has a temporal origin. Since the consideration of such evidence commands the bulk of our essay, we have little space to defend our implicit presupposition of time's reality. One could undercut the evidence for a temporal origin of the universe by denying that presupposition. One might base an "atemporalist" position on the idea that a successful theory of quantum gravity will cease to regard time as a fundamental characteristic of nature [Kiefer, 2008]. Here we offer the following brief comments on the atemporalist position:

1. *The atemporalist position is uncritically literalist with regard to model interpretation*. Our interest is in which cosmogonic model most nearly approximates reality. The mere existence of a useful mathematical formalism does not imply that the entities in that theory correspond with reality. The past (and present) usefulness of GTR, for example, never meant that we were obligated to make an ontological commitment to its view that gravity *just is* the curvature of an objectively real spacetime. Nor are we obligated to believe that time does not exist because a "pure" Wheeler-DeWitt formalism does not contain it and the WKB surrogate is described as merely an "approximation." The Wheeler-DeWitt quantum gravity approach describes both the disappearance of absolute time that results from the attempt to quantize Einstein's gravity as well as time's reappearance as a semiclassical approximation. The fact that time is recoverable as an approximation, however, does not imply that there is either an elimination of time ("pure" Wheeler DeWitt) or that the superposition of all possible metrics necessarily

requires an indeterminate meaning to time *in reality*. Useful models exist at different levels of abstraction from physical reality. We therefore feel perfectly comfortable invoking the approach of Kiefer (and his sometime colleague H. D. Zeh) in our discussion of cyclic universes and a reversed arrow of time (see section 4.6) on the grounds that one need not attach ontological commitment to their larger metaphysical claims regarding the elimination of time.

The issue at hand is therefore not the narrow issue of whether there are quantum gravity modeling schemes that are atemporalist. Rather the question is what determines our ontological commitments. Although contemporary philosophy of time has been deeply imbued with the naturalized epistemology of W. V. O. Quine [1969], which shuns so-called "first philosophy" in favor of taking the deliverances of the natural sciences as authoritative guides to reality, such epistemological naturalism is by its very nature simply a methodological disposition, which cannot itself be grounded either scientifically or in any other way, to restrict one's basic sources of evidence to the natural sciences [Rea, 2002, pp. 63–67; *cf*. pp. 1–7]. Not being so disposed, we see no reason that physics should subvert the veridicality of our experience of tense and temporal becoming, much less of time itself. Indeed, even the illusion of temporal becoming entails the reality of temporal becoming in the contents of consciousness. We find it odd that it is "ordinary" quantum mechanics which brings to the table a near Newtonian view of absolute time, yet this is the ingredient that supposedly spoils the notion of temporality. It is "ordinary" quantum mechanics that enjoys the impressive experimental verification. We should take care in transferring a borrowed legitimacy to a quantum gravity theory that itself lacks consensus.

It seems to us appropriate to consider each and every scheme that cosmologists bring to the table to throw light on the question of which models approximate physical reality. We are not wedded to the specific concept of metric time found in GTR. Authors of cosmogonies routinely attempt to address either the origins question itself or, at least, questions regarding the universe's initial conditions. Many quantum cosmogonies include claims of past eternality (and thus presuppose the reality of time). Others discuss cosmic beginnings *ex nihilo*. Surely the cosmologists who formulate these theories think that they are doing work relevant to the actual universe and not merely mathematical exercises that have no basis in reality. Hence, it seems to us a profitable and important exercise to address such models.

2. *Quantum Gravity models must replicate the observed behavior that more primitive models of the universe (read GTR) seek to model*. Isaac Asimov, in an article entitled "The Relativity of Wrong" [Asimov, 1989], pointed out the fallacy of believing that current "primitive" models lack usefulness because they are "wrong," given that a future theory is always just over the horizon to overturn the current one. There are degrees of "wrongness," and theories which are less wrong than their predecessors are so because they take account of what is empirically correct in their forebears. It is simply mistaken to think that phenomena predicted and analyzed by GTR (especially indirectly *observable* phenomena such as black holes) via such things as singularity theorems cease to be relevant because a different mathematical formalism may

someday come into general use. There may be no such things as singularities, *per se*, in a future quantum gravity formalism, but the phenomena that GTR incompletely strives to describe must nonetheless be handled by the refined formalism, if that formalism has the ambition of describing our universe. This can be seen in the above example of the Wheeler-Dewitt approach as described by Kiefer. Big Bang Friedmann universes with a meaningful time coordinate are recoverable in the theory (as Kiefer shows) but are described as a semi-classical WKB approximation which invokes a mechanism called decoherence. They, in fact, must be recoverable because that is what we *see*. If Big Bang Friedmann-like behavior were not recoverable, that in itself would be grounds for rejecting the larger theory.

3. *Preservation of time as a fundamental property of the universe is possible within a quantum gravity approach*. As examples, see [Mersini-Houghton, 2009], [Smolin, 2009], [Sorkin and Rideout, 2004], [Carroll, 2008], and [Loll, 2008]. Mersini-Houghton defends an arrow-less but nonetheless global and beginningless temporal axis within a larger multiverse. Carroll wants a global time with an arrow that points toward greater entropy given any arbitrary point of departure within the multiverse. Smolin defends a similar view in his postulation of multiple universes birthed in a temporal sequence through black hole incubators. His article "The Unique Universe" is a full fledged attempt at refutation of a timeless multiverse. Yet his defense of the concept of fundamental time is *not* a defense of GTR's metric time. Rather it defends temporality within the context of *background independent* (*i.e.*, emergent spacetime) quantum gravity theories. An example of this approach is the work of Renate Loll which we briefly discuss in section 5.0. Another approach which includes explicitly the property of tense and at least half of the doctrine of presentism (*i.e.*, the thesis that the future does not exist) is the "causal set theory" of Rafael Sorkin, David Rideout, and other prominent theorists, such as Faye Dowker.

In our view, these approaches have a leg up on timeless multiverse quantum gravity approaches precisely because they are consistent with the veridicality of our experience of time and tense. In order to justify a claim so extraordinary as the unreality of time, evolution, and illusoriness of our experience thereof, the timeless multiverse needs something beyond demonstration as a 'consistent' quantum gravity formalism.

4. *The atemporalist position agrees with ours with regard to the fact that different universes within a multiverse are now incommensurate with respect to a temporal measure*. Lastly, the atemporalist position may actually prove to be supportive of the fact of the finitude of the past, if its attempt to eliminate time as a measure should turn out to succeed only in part. Suppose, for example, that the concept of "emergent time" turns out to be meaningful along with the objectivity of tense. (We discuss this in section 4.6.) The impact on supposed "pre-Big Bang" cosmogonies can be enormous, if one can no longer stretch a meaningful time back "through" the Big Bang "singularity." If universe phase *A* is incommensurate with respect to time with universe phase *B*, then it is just false to suggest that phase *A* precedes phase *B*. If we suppose that time is a meaningful local concept within *A* and *B*, then the implication is that both *A* and *B* have temporal origins. This is supportive of the finitude of the past and, hence, of the beginning of the universe.

4.2 Expanding universe

Pitts seems to think that the person who believes that current physical cosmogony lends significant support to (2) must be what he calls a "GTR exceptionalist," that is, someone who tolerates singularities (and, perhaps, takes GTR to be the fundamental theory of gravity). A GTR exceptionalist would be a person who holds that our universe obeys the Hawking-Penrose singularity theorem [Hawking and Penrose, 1970]. Hawking-Penrose assumed that gravity is always attractive (that is to say, their theorem dealt only with "ordinary" types of matter). Just about 30 years ago, however, the idea of "inflation" was introduced into cosmology as a resolution to several of the anomalies in the standard hot Big Bang picture, and a key characteristic of inflation is that it is driven by a type of energy that violates the energy condition of the Hawking-Penrose theorem. Thus it became a live question as to whether inflationary cosmologies might accommodate a beginningless past. Perhaps we are living in a vacuum bubble that resides in a much larger (perhaps past and future fractal) structure of nested bubbles.

On this view, a small volume of "false vacuum" may decay into the "true vacuum" in which we live. This process, first suggested by Sidney Coleman and Frank De Luccia [Coleman and De Luccia, 1980] would identify our Big Bang with one of these quantum tunneling "decay" events. Thus the Big Bang would not represent an absolute beginning of space and time. Rather there would be a pre-Big Bang past.

What Pitts does not mention, however, is that singularity theorems have also evolved in the intervening 30 years. There has been a lively debate centered (mostly) around the role of energy conditions in these theorems. In 2003 Arvind Borde, Alan Guth, and Alexander Vilenkin published a singularity theorem that was completely independent of energy conditions. Vilenkin explains:

> A remarkable thing about this theorem is its sweeping generality. We made no assumptions about the material content of the universe. We did not even assume that gravity is described by Einstein's equations. So, if Einstein's gravity requires some modification, our conclusion will still hold. The only assumption that we made was that the expansion rate of the universe never gets below some nonzero value, no matter how small. This assumption should certainly be satisfied in the inflating false vacuum. The conclusion is that past-eternal inflation without a beginning is impossible [Vilenkin, 2006, p. 175].

In their formal paper, Borde, Guth, and Vilenkin state:

> Our argument shows that null and time-like geodesics are, in general, past-incomplete in inflationary models, whether or not energy conditions hold, provided only that the averaged expansion condition $H_{av} > 0$ holds along these past-directed geodesics [Borde, Guth, and Vilenkin, 2003, p. 3].

Borde et al. suggest that timelike and null geodesics are redshifted in energy as they are stretched out in an expanding space. Looking backwards in time, they are blueshifted.

The blueshift becomes infinite in a finite amount of proper time (or affine parameter for null geodesics). In GTR, the infinite blueshift suggests a singular condition.

The BGV conclusion applies beyond the inflationary model class. Any universe (including universes modeled by higher dimensional cosmology, pre-Big Bang cosmology, and so forth) which, on average, expands has to connect, in a finite time, to a past boundary.[11] According to Vilenkin, this result is independent of modifications to Einstein's equations (which could be the result of "low energy" corrections coming from a quantum gravity approach) or the particular characteristics of the proposed cosmogony. If the universe (or multiverse) expands (on average), then it has a beginning, period.

It is possible to object that the BGV theorem is a classical result lacking the rigor of a full quantum gravity approach. The objection is that BGV is based on metric tensors, which one could claim would not exist in a "full" theory of quantum gravity.

As discussed in section 4.1, we believe the objection is without merit. We add that Borde, Guth, and Vilenkin are not silent on the topic of quantum gravity. They indicate that their results demonstrate a boundary where a quantum gravity theory will be necessary to describe further the physics. Thus the classical dynamics governed by their theorem are adequate to describe the universe at "low" energy states, including its past history roughly back to the Planck time. There is then a quantum gravity reckoning:

> What can lie beyond this boundary? Several possibilities have been discussed, one being that the boundary of the inflating region corresponds to the beginning of the Universe in a quantum nucleation event. The boundary is then a closed spacelike hypersurface which can be determined from the appropriate instanton.
>
> Whatever the possibilities for the boundary, it is clear that unless the averaged expansion condition can somehow be avoided for all past-directed geodesics, inflation alone is not sufficient to provide a complete description of the Universe, and some new physics is necessary in order to determine the correct conditions at the boundary. This is the chief result of our paper [Borde, Guth, and Vilenkin, 2003, p. 4].

As briefly mentioned earlier, a Feynman path integral based cosmogony such as the Hartle-Hawking 'no-boundary' [Hawking & Hartle 1983], or the Vilenkin 'tunneling from nothing' approach [Vilenkin 1982] can be interpreted as a creation-ex-nihilo[12] that would explain the boundary suggested by the BGV theorem as well as lend support to *kalām's* second premise.

We discuss this quantum reckoning in section 5 of our paper. So either the boundary represents a quantum nucleation event or else (within the context of Pitts' original objection) we are left with the options discussed in sections 4.3–4.6 of this paper.

A possible further objection to the BGV theorem might draw upon Pitts' suggestion [Pitts 2009, p. 681] that the Bach-Weyl approach to GTR does not assign lengths to curves. Demonstrating a past inextendible geodesic (*i.e.*, one of finite length) is precisely how BGV purport to show that their boundary exists. Does the Bach-Weyl formalism give us, then, reason to doubt that objective scales exist?

We question whether Bach-Weyl should be taken as a realist approach to modeling. It may well be better understood instrumentally.[13] We note that some theorists seem to prefer a more modest use of conformal invariance. Roger Penrose's Cyclic Conformal Cosmogony (see section 4.5), for example, uses exactly the Bach-Weyl formalism to model the boundary between consecutive oscillations. But Bach-Weyl requires all contents of the universe to be conformally invariant. Massless particles such as the photon and the graviton meet this criterion. In Bach-Weyl one splits a metric into a scale-free part and another part that measures space and time intervals. But massless particles (only) do not experience a passage of time. So if the universe contained only massless particles, one could argue (as Pitts does) that one could drop the part of the metric that represents intervals. But where *mass exists*, conformal invariance does not. So for all phases of the universe where mass exists (which in the real universe is pretty much every era after the Big Bang), one needs the full metric. Penrose thus splits his model into two epochs: one consistent with conformal invariance and one without it.

Similarly Gerard't Hooft, an advocate of the Bach-Weyl approach, describes other necessary parts of a cosmogony that are not scale invariant and are thus problematic for a "pure" Bach-Weyl methodology. In a paper defending the existence of spacetime (as an objective feature of nature) at the expense of "emergent time" theories he states:

> Describing matter in a $\hat{g}_{\mu\nu}$ metric will still be possible as long as we restrict ourselves to conformally invariant field theories, which may perhaps be not such a bad constraint when describing physics at the Planck scale. Of course that leaves us the question where Nature's mass terms come from, *but an even more urgent problem is to find the equations for the gravitational field itself, considering the fact that Newton's constant* GN *is not scale-invariant at all. The Einstein-Hilbert action is not scale-invariant.* Here, we cannot use the Riemann curvature or its Ricci components, but the Weyl component is independent of ω, so that may somehow have to be used. ['t Hooft 2009, p. 8, emphasis added]

't Hooft ultimately believes that scales are introduced in nature via the fundamental processes of information transfer. He sums up his paper by arguing, "The density of this information flow may well define the Planck length locally, and with that all scales in Nature." Thus he supports the idea that scales (time and space intervals) are objectively real features of nature.

It therefore seems to us that the mere fact of the Bach-Weyl approach (which is all that Pitts offers) proves little. It may describe possible worlds that are fully conformally invariant, but we are interested in a model that can represent the matter-filled universe we see around us.

What, then, has been the reception of the BGV theorem by the physics community? It, as well as its implications, is largely uncontroversial.[14] Instead, a new round of model building has ensued based on possible exceptions to this theorem. Four alternatives present themselves:

1) Infinitely contracting universe "bouncing" into an expansion phase (average expansion < 0; example: de Sitter cosmogony)
2) Asymptotically static universe (average expansion = 0; example: Emergent model class)
3) Eternally cyclic universe (average expansion = 0; example: Penrose conformal cyclic cosmogony)
4) Reversal of the arrow of time (example: Aguirre-Gratton model)

Leaving (4) aside for the moment, let us consider universes that do not, on average, expand over their past histories.

4.3 Infinitely contracting universe

Is a pre-Big Bang infinite contraction (followed by a rebound at a near singular condition into our present-day expansion) as envisioned in (1) a promising alternative? We suggest not. George Ellis comments on the problems that bedevil such an approach:

> The problems are related: first, initial conditions have to be set in an extremely special way at the start of the collapse phase in order that it is a Robertson-Walker universe collapsing; and these conditions have to be set in an acausal way (in the infinite past). It is possible, but a great deal of inexplicable fine tuning is taking place: how does the matter in widely separated causally disconnected places at the start of the universe know how to correlate its motions (and densities) so that they will come together correctly in a spatially homogeneous way in the future??
>
> Secondly, if one gets that right, the collapse phase is unstable, with perturbations increasing rapidly, so only a very fine-tuned collapse phase remains close to Robertson-Walker even if it started off so, and will be able to turn around as a whole (in general many black holes will form locally and collapse to a singularity).
>
> So, yes, it is possible, but who focused the collapse so well that it turns around nicely? (pers. comm., January 25, 2006)

First, then, such models encounter the significant problem of acausal fine-tuning. One asserts not just brute contingency but also a very curious form of it. In the face of apparent fine-tuning, physicists usually prefer to offer some type of explanation. For example, cosmologists are avidly seeking an explanation for apparent fine tuning of the parameters of the standard model (such as force coupling constants) in the form of a multiverse or a superdeterministic Theory of Everything. Or one thinks of Guth's inflationary resolution of the horizon problem (past thermodynamic equilibrium). If we are going to give up explanation, then what was wrong with leaving cosmology as it was prior to 1980, namely, the standard hot Big Bang model (with the associated breakdown of physics at the singularity)?[15]

The second problem is that the collapse becomes chaotic as it approaches the singularity. This will produce a pre-expansion start condition that is known to be dramatically different from our actual Big Bang. This phenomenon is referred to as

"BKL chaos" after its discoverers [Belinsky, Khalatnikov, and Lifshitz, 1970]. The same problem will appear for all attempts at a past-eternal timeline that seek to introduce a pre-Big Bang phase that "bounces" into the present expansion. In fact, the real implication of BKL may well be that it is physically impossible to "bounce" *through* a "singularity."[16] So option (1) is unpromising.

4.4 Asymptotically static universe

An asymptotically static space, as envisioned in option (2), is one in which the average expansion rate of the universe over its history is equal to zero, since the expansion rate of the universe "at" past infinity is zero (thus "infinity" dominates any finite expansion phase, no matter how long). Hence, the universe, perhaps in the asymptotic past, is in a static state (neither expanding nor contracting). This feature of the model allows it to escape the BGV singularity theorem. Consider, for example, the GTR-based "Emergent" model class of George Ellis *et al.* [Ellis, Murugan, and Tsagas 2004], [Ellis and Maartens 2004]. It features two stages; an Einstein Static State (ESS) and an inflationary phase that leads to our present, dynamic, expanding universe.

The Einstein static universe itself was originally viewed as past eternal. But there are obvious problems with this model. The reason Einstein himself dropped it was its feature of unstable equilibrium. Although, in pure non-quantum GTR, one can consider a static state with worldlines that trace to negative infinite time, in reality we know that gravity is a quantum force. As Vilenkin notes, "Small fluctuations in the size of the universe are inevitable according to the quantum theory, and thus Einstein's universe cannot remain in balance for an infinite time" [Vilenkin, 2006, p. 209]. On the other hand, the current observable universe is demonstrably *not* in a static state. So there would need to be at least two stages in such a model, a primordial ESS followed by an expansion phase. A quantum fluctuation would force a transition from ESS to an expanding universe. But this very mechanism implies that the *initial state* is not past eternal, since such a fluctuation will inevitably occur within a finite time.[17]

To preserve past eternality, one could claim that the static state is only an "ideal point" asymptotically approached at past infinity. Some philosophers have expressed reservations, however, over the contrived nature of the past infinity featured in such an interpretation. The cosmological (and thermodynamic) arrows of time are so weak that they are indistinguishable from a "timeless state." For example, Rüdiger Vaas, characterizing the Emergent models as "soft-bang/pseudo-beginning" in nature, views the asymptotic approach toward ESS as a mathematical artifact [Vaas, 2004, p. 18].

A further problem with these models is the implied fine tuning of the initial state. One way to look at this is to imagine starting in the present and extrapolating the past history of the universe. Will the resultant evolution, with high probability, produce an ESS state? No, such an outcome is improbable. Ellis is sensitive to the fine-tuning problem but thinks his approach worth pursuing in the absence of a quantum gravitational resolution [Ellis and Maartens, 2004, p. 228].

Subsequent to his initial work on Emergent models, which were based solely on GTR, Ellis along with his colleagues integrated a quantum gravity approach

into their models. Ellis' colleague David Mulryne elaborates: "The importance of the [equilibrium state as described by quantum gravity] is that, in contrast to the [ESS] solution present in GTR, slight perturbations do not result in an exponential divergence from the static universe, but lead instead to oscillations about it" [Mulryne *et al*., 2005, p. 6]. The ESS is now viewed as a "low-energy" solution of loop quantum gravity (LQG) to make the Einstein state stable against perturbations of a limited size. LQG theorist Martin Bojowald explains how Mulryne *et al*. use the mechanism of perturbation to cause an initially stable oscillation to escape to full blown inflation:

> Static solutions do not evolve, and so are clearly ill-suited as a model for the Universe. *But by introducing a perturbation to a static solution, one can slightly change it and thereby start a more interesting history*. Unfortunately, the classical solution [ESS] is unstable: any disturbance grows rapidly, leaving little of the initial state behind. The insight of Mulryne and colleagues is that quantum effects could supply all the necessary ingredients where classical solutions do not. Within the framework of loop quantum gravity, repulsion also implies static solutions at small size, but these – in contrast to the classical case – are stable. *According to the authors' model, perturbing such a state leads to small cycles of interchanging expansion and contraction*. During this process, matter will evolve slowly, and the cycles will gradually change their behavior. By itself, this perpetual recurrence and incremental change seems to lack the spark necessary for so momentous an event as the birth of the Universe. And indeed, Mulryne and colleagues identify one final theoretical ingredient that lights this spark: mediated through repulsive effects, *potential energy is gradually pushed into the matter during its slow evolution. At the point when potential energy starts to dominate kinetic energy, the mundane cycling is broken by a sudden, dramatic inflationary explosion – the emergent Universe* [Bojowald 2005, pp. 920–1; our emphasis].

Should the past be judged infinite in such an approach? There *is* a claim by its authors to possible past eternality: "The universe undergoes a series of non-singular oscillations in a (possibly) past-eternal phase with the field evolving monotonically along the potential" [Mulryne *et al*. 2005, p. 6].

Similar to the earlier Emergent models of Ellis *et al.*, which proposed an ideal ESS point "reached" in the infinite past, the model of Mulryne *et al*. proposes that any point which is close to the stable equilibrium point could serve as an "initial" state for the model. The size of the universe, then, oscillates about this initial state, always maintaining a non-zero positive value. This oscillation has an amplitude which is associated with a maximum size for the universe (per oscillation) which we will call $\mathbf{amax}_i$ (see [Bojowald and Tavakol 2008a, figure 2] and [Mulryne et al. 2005, figure 7]). In the absence of the self-interaction property of the scalar field φ, this condition is said to be perpetual. But the potential energy in the system must grow, forcing **amax** to grow with each subsequent cycle. This is said to introduce a cosmological arrow of time, which is based on the evolutionary increase in **amax**. We will take the equation

amax(t) to be fundamental; the potential V(φ) and scalar field φ(t) are reverse engineered to produce it; see [Ellis, Murugan, and Tsagas 2004, section II text and figure 1], [Mulryne *et al.* 2005, figure 5].

In the limit t → -∞, it is desired that the change in **amax** (call it $\Delta\mathbf{a}_{\mathbf{evolution}}$) between cycles asymptotes to zero. Note that the interval in time between a cycle with maximum amplitude $\mathbf{amax}_i + \Delta\mathbf{a}$ (with arbitrarily small Δa) and the present is finite; this is shown by the numerical analysis in [Mulryne *et al.* 2005]. So if the initial process of reaching $\mathbf{amax}_i + \Delta\mathbf{a}$ has a finite timeline, the full timeline to the present is finite.

Now, consider a quantum fluctuation of the scale factor, $\Delta\mathbf{a}_{\mathbf{fluctuation}}$. This is a separate, instantaneous, mechanism whereby the universe may reach a larger size. For the universe at **amax** [Bojowald and Tavakol 2008a, figure 2 and equations 32 and 33], and in the past asymptotic limit of an unsqueezed quantum state,[18] the fluctuation is roughly proportional to **amax**. So looking to the infinite past, $\Delta\mathbf{a}_{\mathbf{fluctuation}}$ will be >> $\Delta\mathbf{a}_{\mathbf{evolution}}$.

This suggests two things. First, as one looks to the past, the (typical) size of fluctuations of the scale factor eventually exceeds and dominates the evolutionary increase (in **amax**) between cycles. As Vaas might describe it, there is no cosmological arrow of time in this region. Second, it does not seem reasonable to assert that the universe will need an infinite amount of time to grow $\mathbf{amax}_i$ to amplitude $\mathbf{amax}_i + \Delta\mathbf{a}$ (where Δ**a** can be arbitrarily small). The probability should be unity that it will reach that goal by instantaneous, discrete transition in finite time, as opposed to continuously evolutionary growth over infinite time.

Furthermore, since the early phase of this model features an oscillating universe, it also encounters an entropy objection to past eternality (see section 4.5). It seems, then, that the past eternal nature of this model is a mathematical artifact. One need not reject the model; one merely notes that there must a temporal origin of some sort in the finite past.

In sum, here is a counterexample to Pitts' notion that a past lacking a singularity must of necessity be eternal.[19] The Emergent models that feature (only) a GTR based ESS are metastable.[20] An improvement to the Emergent model resolves this instability by providing quantum-based modifications to the Friedmann equation. As the authors say, this provides "partial" ameliorization of the metastability objection and the fine-tuning objection [Mulryne *et al.*, 2005, abstract]. But the nature of quantum fluctuations demands that one cannot have an infinite timeline during which an arbitrarily small, continuous evolution in scale factor occurs. It seems to us this objection would apply, generally, to asymptotically static constructions.[21]

So far, then, we suggest that there are good arguments to believe that contemporary pre-Big Bang universe models that contract, are static, or expand (on average) either imply a beginning or are untenable. The expanding models obey a newly discovered singularity theorem (BGV) and likely have a beginning of a quantum nature. The contracting models featured in alternative (1) seem to be unpromising and add nothing to the standard hot Big Bang model from the standpoint of *explanation*. The static models featured in alternative (2) imply a beginning but do not necessarily describe how that state came to be.

4.5 Eternally cyclic universe

What, then, about cyclic models, as envisioned in option (3) above? According to these models, the universe goes through a cycle in which it grows from zero (or near-zero) size to a maximum and then contracts back to its starting condition. The universe itself is periodic, in the sense that it undergoes many such cycles, perhaps an infinite number. The average expansion of the universe would be zero in a "pure" cyclic model, since cycle by cycle, the universe always undergoes precisely equal amounts of expansion and contraction. Hence, a cyclic model evades the BGV Theorem.

Cyclic models face a well-known thermodynamic problem. As Vilenkin notes, "A truly cyclic universe has a problem with entropy increase: it should have reached thermodynamic equilibrium by now" [pers. comm., January 19, 2008]. Our observation of the present-day universe indicates that we are not at a condition of thermodynamic equilibrium – a good thing for us, as life requires nonequilibrium conditions in order to exist. As one looks into the past, the size of each cycle is also thought to decrease (due to radiation effect on entropy; see [Barrow and Dabrowski, 1995]). Eventually a "Planck entropy" could be reached (the minimum currency of entropy exchange), which would preclude the existence of still earlier cycles. A problem similar to that encountered for an asymptotically static cosmogony can also occur. *If one is attempting a realist ontology*, the backward evolution of a cyclic model must stop once quantum fluctuations in volume are of the same size as the volume expectation value. One can continue pursuit of the problem (the classical background disappears) in a full quantum approach (pers. comm. With Martin Bojowald, 04–15-2010). If this means that time 'disappears' or that the arrow of time reverses, the implications are discussed in section 4.6.

Cosmologists Thomas Banks and Willy Fischler contend that a contracting space filled with quantum fields will have an "ergodic" property as the space shrinks. Its fields become highly excited as one approaches the end of contraction, and these fields will produce chaotic fluctuations. Spontaneously created matter with a different equation of state will dominate the energy density. That, and the inhomogeneity of the fluctuations, will prevent cycling. Banks and Fischler even suggest that the fields will spontaneously produce a dense "fluid" of black holes leading to a condition they call a "Black Crunch" [Banks and Fischler, 2002] for arbitrary states approaching full contraction.[22]

Martin Bojowald and collaborators [Bojowald, Maartens, and Singh, 2004], [Bojowald 2006], [Bojowald and Tavakol 2008a, 2008b] have done a significant amount of work on building contemporary cyclic models to address these difficulties. The LQG approach does seem to resolve the singularity admirably in its initial tests against simple (homogeneous) models [Ashtekar, Pawlowski, and Singh 2006], [Bojowald 2007], although it is unclear if this approach solves all the problems brought up by opponents such as Banks and Fischler.[23]

In 2006, Bojowald suggested the possibility that the recollapse phase of each oscillation was entropy-reducing (see [Craig and Sinclair 2009, p. 172]). Perhaps the universe does not cumulatively gain entropy cycle to cycle. But if so, this could lead to a different problem in that it is possible to interpret an *entropy reducing* recollapse

in terms of a reversed arrow of time; thus the "recollapse" is really another expanding universe. We shall take up this question again in section 4.6.

In 2008, Martin Bojowald and Reza Tavakol entertained the opposite suggestion; entropy is monotonically gained from cycle to cycle. Quantum squeezing of the conjugate variable pair V (volume) and P (momentum) of the universe was identified as an entropy contributor without classical parallel. Interestingly, the quantum gravity model now displays the same behavior with respect to entropy gain that [Barrow and Dabrowski, 1995] note for classical models in their paper:

> An interesting question is whether in a cyclic model one generically expects to have a finite or an infinite number of past cycles. The problem with the finite case is that it does not resolve the origin question. In the emergent scenarios, as well as some other such models, the universe is assumed to have undergone an infinite number of past cycles so as to remove the question of the origin. In that case any given cycle would have an infinite number of precursors and generically we therefore have to expect the current state to be squeezed ….. The question then is how the squeezing in a generic cycle is determined. *If each cycle produces the same amount of squeezing, a generic cycle would have infinitely squeezed states, which could not be semiclassical ... For growing cycles, as in the emergent scenario, the change in squeezing is initially small and approaches zero for cycles in the infinitely distant past.* Depending on the precise scenario, the sum of all squeezing contributions may converge, such that a finite value results for a generic cycle. [Bojowald and Tavakol 2008a, p. 8, our emphasis]

Similarly Aguirre comments:

> For an observer along that worldline to perceive an arrow of time (AOT) it must see local net entropy generation. To the past, it seems entropy destruction. In a finite neighborhood, there is finite entropy, so going far enough 'to the past' the entropy must (a) asymptote to a constant, (b) start to increase again, or (c) become ill-defined as a singularity is encountered. [Aguirre 2007, p. 30]

So avoiding a reversal of the AOT requires entropy to asymptote to a constant in the infinite past. In a quantum model, the squeezing of states is going to be a factor in accomplishing this. In their paper Bojowald and Tavakol suggest both that a special "unsqueezed" state represents the asymptote and that it is problematic to assert that such a state existed. Bojowald [pers. comm. April 15, 2010] has indicated to us that there were further problems since discovered with his and Tavakol's assumptions on monotonicity of variables. So, as of this writing, the problem of structuring entropy to be infinitely subdividable remains.

Returning to the general discussion; suppose that total entropy can be kept constant with cumulative cycles and that contracting phases are not reinterpreted as having a reversed arrow of time. There remains the issue of dark energy, which may have the potential to halt the cycling and induce an open-ended expansion. The current

empirically observed dark energy effect, for example, appears adequate to produce an open-ended, accelerated expansion. This result would be definitive if the dark energy is of the form of a cosmological constant (that is to say, if its value is independent of space and time; see [Barrow and Dabrowski, 1995]). Indeed, open-ended, accelerating expansion does appear to be the fate of the present-day universe [Overbye, 2006]. If an entropy gain (cycle to cycle) is denied, then the amplitude of the oscillation (maximum scale factor experienced by the universe) is constant. But then one could never have more than one "cycle," for the cosmological constant would lead to open-ended expansion *the first time*. Hence, the initial cosmological singularity (our Big Bang) *would* represent an absolute beginning.

There are some genuinely exotic cyclic models that purport to escape these problems (see [Craig and Sinclair, 2009] for an analysis of various cyclic models current in contemporary cosmological discussions). A fascinating attempt to counter the entropy and dark energy problems mentioned above is Roger Penrose's Cyclic Conformal Cosmogony [Penrose, 2006]. Penrose suggests a solution to the entropy problem by saying that the initial "singularity" is the *same thing* as the open-ended de Sitter-like expansion that our universe seems about to experience. Their mathematical equivalence is demonstrated through an appropriate conformal transformation of one state into another. Penrose explains how we are to think about this situation:

> Physically, we may think that again in the very remote future, the universe 'forgets' time in the sense that there is no way to build a clock with just conformally invariant material. This is related to the fact that massless particles, in relativity theory, do not experience any passage of time ..... With conformal invariance both in the remote future and at the Big-Bang origin, we can try to argue that the two situations are *physically identical*, so the remote future of one phase of the universe becomes the Big Bang of the next [Penrose, 2006, p. 2761].

Penrose admits that his view is heterodox in that, for this envisioned scenario to obtain, all massive fermions and massive, charged particles must disappear to radiation, including, for example, free electrons. He concedes that there is (currently) no justification for positing this.

The Cyclic Conformal Cosmogony is based on the Weyl Curvature Hypothesis (WCH) and Paul Tod's implementation of this idea within GTR [Tod, 2003]. WCH is defined as follows:

> Weyl curvature is the kind of curvature whose effect on matter is of a distorting or tidal nature, rather than the volume-reducing one of material sources ….. The physical conjecture that I refer to as the Weyl curvature hypothesis asserts that (in some appropriate sense) the Weyl curvature is constrained to be zero (or at least very small) at initial singularities, in the actual physical universe [Penrose 2005, pp. 765, 768].

Penrose describes the mathematical technique necessary to stitch a singularity to a maximally extended de Sitter expansion:

> Tod's formulation of WCH is the hypothesis that we can adjoin a (past-spacelike) hypersurface boundary to space-time in which the conformal geometry can be mathematically extended smoothly through it, to the past side of this boundary. This amounts to 'stretching' the metric by a conformal factor Ω which becomes *infinite* at the Big Bang singularity, so that we get a smooth metric $\hat{g}_{ab}$ which actually extends across this boundary. So far, we regard the conformal 'space-time' prior to the Big Bang as a mathematical fiction, introduced solely in order to formulate WCH in a mathematically neat way. However, my 'outrageous' proposal … is to take this mathematical fiction seriously as something *physically real* [Penrose 2006, p. 2761].

A potential failing of this approach is the supposed correspondence between Weyl curvature and entropy. The correspondence seems clear enough when one is considering the structure of the initial Big Bang singularity, given its vanishingly small entropy state. But while the de Sitter-like end state of the universe also minimizes Weyl curvature, its entropy is *maximized*. Like a black hole obeying the Hawking-Bekenstein entropy law, a de Sitter space has a cosmological horizon with entropy proportional to its area. It is generally believed that this state represents the maximum entropy that can fit within the horizon.

Penrose has recently chosen to regard the entropy of the cosmological horizon as spurious and to invoke non-unitary loss of information in black holes in order to equalize the (vanishingly small) entropy at the boundary [Penrose 2009, p. 15]. Penrose attributes the large entropy at late universe times (but before significant decay of the universe's black hole population) to degrees of freedom internal to the black holes. He then suggests that in CCC the universe's entropy is "renormalized," so that we can discount the entropy contribution from the horizon when all black holes have evaporated.

On the other hand, it seems, Penrose notwithstanding, that the entropy of the cosmological horizon *must* have physical meaning, since the entropy of the de Sitter-like system outside a black hole must be higher than that of the black hole itself. For if it were not, then the physics of black hole decay (upon which Penrose's scenario depends) would not work properly. Black hole decay is actually a dynamic system which is the sum of the energy lost by Hawking radiation plus the energy gained by absorption of local matter created by thermal fluctuations due to the de Sitter Gibbons-Hawking temperature. The mere fact of these thermal fluctuations suggests that the entropy of the cosmological horizon is a real physical manifestation as opposed to an accounting gimmick.

As a thought experiment, one can imagine what happens as more and more external matter (over and above that added by external environment thermal fluctuations) is added to a black hole. Is there a maximum size? As the black hole gets larger, its temperature drops, approaching that of the cosmic background. In fact, at a certain point the black hole itself becomes physically indistinguishable from a cosmic horizon and can be identified as such. At no point does it become unreal. It therefore seems unwarranted to embrace Penrose's latest (2009) position, and very

few physicists have been persuaded by Penrose's non-unitary brand of quantum physics.

In sum, while Weyl curvature is the same between the two states that Penrose wishes to say are identical, the entropy is not. It seems to us, then, that the two states cannot be identical. Hence, we question CCC's viability.

It seems, then, that cyclic models do not avert the conclusion that time had a beginning. Thermodynamic objections to cyclic models, "classical" anticipations of which were seen in Richard Tolman's early work in the 1930s [Tolman, 1934], still seem to find purchase. Entropy should increase from cycle to cycle, implying that (1) we should now be in thermodynamic "heat death" (which we are not), and (2) the maximum size (scale factor) of the universe should grow with each cycle. This large scale "classical" behavior likely ends when (looking into the past) the cycle size shrinks to Planckian physics. If entropy gain is denied, then the empirical observation of a cosmological constant of sufficient size to lead to an open-ended expansion rather than a future contraction (this does seem to be the fate of our universe) *also* implies that there can have been no previous cycle. If our future does include a return to a singular condition, then, as Banks asserts (recalling our earlier discussion of BKL chaos), "It seems silly to imagine that, even if [a future singular condition for the universe] is followed by a re-expansion, that one would start that expansion with a low entropy initial state, or that one had any control over the initial state at all" [pers. comm. Oct 17, 2008].

4.6 Reversing time's arrow

There is a longstanding debate in cosmological circles as to how to interpret universe phases that would be entropy-reducing. There is no necessity (in GTR) for the cosmological and thermodynamic arrows of time to align with each another, although they seem to do so in the present era of our universe. This debate has consequences for discussion of a cosmic origin. For example, in inflationary models, the BGV theorem seems to bar a beginningless past. But suppose one takes a model that features a contraction, a bounce through a near-singular state, and then an expansion. Such a model faces fine-tuning objections such as that discussed in section 4.1. Yet it is possible to (re)consider the contracting phase as having a reversed arrow of time. The boundary that formerly represented the "bounce" will now bisect two symmetric, expanding universes on either side.

On such a view spacetime is extendible beyond the boundary for an observer, looking backwards in time, from either side. Does this imply past eternality? Cosmologists Anthony Aguirre and Steve Gratton think so. They have employed the concept to create an "eternal steady state inflation model" [Aguirre and Gratton 2002]. To visualize their idea, start with the standard foliation of de Sitter space as looking something like an hourglass. The width of the glass represents the scale factor of the universe. Vertical height (pointed upward) is the time. Thus the space shrinks to a minimum radius and then rebounds into an expansion. On this understanding, one cannot possibly have "past eternal inflation." Borde, Guth, and Vilenkin explain:

> The intuitive reason why de Sitter inflation cannot be past-eternal is that, in the full de Sitter space, exponential expansion is preceded by exponential contraction. Such a contracting phase is not part of standard inflationary models, and does not appear to be consistent with the physics of inflation. If thermalized regions were able to form all the way to past infinity in the contracting spacetime, the whole universe would have been thermalized before inflationary expansion could begin [Borde, Guth, and Vilenkin, 2003, p. 1].

To obtain a different outcome, Aguirre and Gratton choose to interpret de Sitter space according to a procedure first suggested by Erwin Schrödinger in 1956. Utilizing "elliptic de Sitter space," they remap the hourglass by splitting it in two along a line from the top left to the bottom right. For all intents and purposes, these are two separate universes. Maulik Parikh explains how this looks to antipodal observers on opposite sides of the divide:

> The antipodal map, $X^I \rightarrow -X^I$, [where 'I' is a dimensional index] changes the sign of the time coordinate of the embedding space, and also that of the direction of time in de Sitter space. The resulting quotient space, dS/Z2, *is as a consequence not time-orientable*: although one can locally distinguish past and future, *there is no global direction of time*. This fact clearly changes many standard notions about space and time that we are accustomed to. For instance, it is impossible to choose a Cauchy surface for elliptic de Sitter space that divides spacetime into a future and a past region [Parikh *et al.*, 2002, p. 6, emphasis added].

Aguirre and Gratton explain the relationship between the two disconnected regions:

> In essence, this construction partitions the full de Sitter space-time into a self-consistent set of two noncommunicating [steady state] universes. An observer in region I does not see anything in its past light cone from an observer in region II because that other observer cannot signal into its past, and vice-versa. *Seen in this way the boundary condition forbidding physical particles from following geodesics across [the boundary] – into one universe is in no way strange or unreasonable, as it follows directly from the forbidding of causality violations in the other universe.* (One could similarly partition de Sitter space-time by any non-timelike boundary B away from which time flows).
>
> ….. Without the identification, the space-time manifold is time-orientable in the mathematical sense that it is possible to continuously divide non-spacelike vectors into two classes which can be labeled 'future' and 'past.' In our construction these labels will only correspond to the physical [arrow of time] in one of the two regions. With the identification the space-time is still a manifold but is not mathematically time-orientable. The physical [arrow of time] is, however, still well-defined and no physical observer will see it reverse [Aguirre and Gratton, 2002, p. 3, emphasis added].

As can be seen, neither region of the remapped space stands in a relation of *earlier than* to the other. There is no global definition of time. As Parikh notes above, it is impossible to determine a future and past region (when considering the whole hourglass). A matter of some importance is that there is no communication; no *causality* by definition from one region to the other. Thus one can say that there is no eternal past that evolved into our present. So Aguirre and Gratton's use of the term "past eternal" seems to be idiosyncratic in that they construe one's ability to draw "complete" geodesics (meaning "from one end of the spacetime to the other," even though no observer or causal chain could traverse them) from end to end on a graph to be a sufficient condition of "eternal." What one really has in this model are two separate universes that trace their origins to a past boundary (the very singularity that Borde, Guth, and Vilenkin have demonstrated).

Aguirre and Gratton conclude: "We suspect that a construction like that proposed here may be necessary in any reasonable model for an eternal universe that avoids a beginning of time" [Aguirre and Gratton, 2002, p. 6]. We suggest, on the contrary, that if they are right, then a beginning of time *has* been demonstrated. Rather than a past eternal *universe*, one has a past finite *multiverse*. This model is a perfectly viable example of a cosmological model that averts the singularity theorems yet still "begins to exist," despite a technical past extendibility through the boundary.

We may resume here our discussion of cyclic models in section 4.5. Martin Bojowald has recently argued that the bounce events of a cyclic model actually represent a scenario similar to the Aguirre-Gratton "double Big Bang" described above. Measurements of a state of the universe "after" a bounce cannot realistically be used to derive knowledge of the state of the universe "before" the bounce. Thus the model contains "cosmic forgetfulness" as a salient feature. Similarly, an observer just "before" a bounce would find that attempts at predicting the post-bounce future suffer from cosmic forgetfulness as well. Bojowald interprets the situation as follows:

> The kind of cosmic forgetfulness realized in this model provides an orientation of time, telling us not only which of the properties before the big bang can be forgotten, but also what direction 'before the big bang' is. An observer after the bounce would be unable to reconstruct the full state before the bounce, but could easily predict the future development toward larger volume. This arrow agrees with the standard notion.
>
> Now asking how an observer before the big bang would experience the same situation, the answer is also clear: such an observer would be unable to determine the precise state at larger values of φ [which is a matter field used in the role of a clock] beyond the bounce, but could easily extrapolate the state to smaller values of φ [the direction away from the 'bounce']. The state at smaller values of φ can be predicted, while the state at large values of φ is forgotten once the bounce is penetrated. *Since one cannot forget the future, such an observer must be attributed a reversed arrow of time*, pointing toward smaller φ. *At the bounce, two arrows would emerge pointing in opposite directions as far as φ is concerned*. In this sense, the model resembles [Aguirre-Gratton and similar suggestions from two other modeling teams using different methods] …..

> Taking the simplest models of loop quantum cosmology at face value is often seen as suggesting the big bang transition to be viewed as a smooth bounce, as one further element not just in a long history of the universe itself but also in a long history of bouncing cosmological models. Some indications, however, suggest otherwise. The bloomy scenario of loop quantum cosmology may well be this: a universe whose time-reversed prehistory we cannot access but which we grasp in the form of initial conditions it provides for our accessible part; a pseudo-beginning; an orphan universe, shown the rear-end by whatever preceded (and possibly created) it [Bojowald 2009, p. 15 (emphasis added].

Recall we had earlier discussed the breakdown of models featuring a classical background plus quantum fluctuations. A hypothetical resolution is to go to a full quantum gravity model. Would this produce past eternality? Bojowald suggests it produces a reversed arrow of time on the other side of the boundary. But, as the moments on one side of the boundary are in no sense earlier than the moments on the other side, there would be no infinite past.

H. D. Zeh, in the same volume in which Bojowald's essay appears, gives independent justification for accepting the interpretation of the boundary as an origin as opposed to a "bounce". The Wheeler-DeWitt (WDW) wave function (upon which loop quantum gravity is based) would be symmetric on both sides of a "bounce." This suggests that what one really has is a "double Big Bang" coming from the same creation event. Zeh also suggests that the semi-classical approximation which defines time in a WDW approach cannot, with validity, be extended "through" a bounce. Lastly, the low entropy initial state is common for all branches of the wave function.

> According to loop quantum cosmology, the Wheeler-DeWitt equation (in this theory replaced by a difference equation with respect to [scale factor] a) can be continued through $a = 0$ to negative values of a. The configuration space of three-geometries is in this way duplicated by letting the volume measure assume negative values (turning space "inside out" while going through $a = 0$). Since the Hamiltonian does not depend on the newly invented sign of a, however, the Wheeler-DeWitt wave function must be expected to be symmetric under this parity transformation, too. Its continuation would then have to be interpreted as an added superposition of other physically *expanding* universes. Since the WKB times, which represent classical time, can *not* be continued through $a = 0$, the interpretation of negative values of a as representing pre-big-bang times is highly questionable. The fundamental arrow, including its consequence of decoherence outside the validity of a WKB approximation, must depend on some low entropy "initial" condition in a for all other ("spacelike") degrees of freedom that occur as physical arguments of the Wheeler-DeWitt wave function. It would be hard to understand how the low entropy state at $a = 0$ could have been "preceded" by an even lower entropy at $a < 0$ in order to avoid a reversal of the thermodynamical arrow in the classical picture of an oscillating universe. [Zeh 2009, pg. 11–12, emphasis in original]

On this view, a loop quantum approach predicts an origin out of which a multiverse arises. Continuation through the "bounce" boundary actually represents other physically expanding universes. Zeh and Claus Kiefer arrive at a similar conclusion arguing directly from the Wheeler-DeWitt equation within the formalism of quantum geometrodynamics.

> One might wonder what happens in the case of models which classically describe bouncing cosmologies: the Universe would then undergo many, perhaps infinite, cycles of expansion and recollapse. What would happen with the entropy in these cases? If the entropy were indeed correlated with the scale factor, as the scenario discussed above suggests, the arrow of time would not continue through a turning point. The bouncing models would thus make no sense in quantum cosmology; one would only have branches of the wave function in which the arrow would point from small to large universe and where time would end when approaching a classical turning point [Kiefer 2009, p. 9].
>
> *The different quasiclassical branches of the wave function which are connected by "quantum scattering" at the turning point should rather be interpreted as all representing different expanding universes*, which disappear at the turning point by means of destructive interference (similar to their coming into existence as separate Everett branches from a symmetric initial state at the big bang) [Zeh and Kiefer, 1994, p. 4152, emphasis added].

Here, Wheeler-DeWitt physics (with a boundary condition of a vanishing wavefunction) seems to predict the creation *ex nihilo* of an Everett multiverse as opposed to a recollapse.

Recalling Penrose's model, suppose that we grant his reversal in entropy as one looks to the universe's far future. One may just as well interpret his model as having a reversed arrow of time when looking to the "past" of a Big Bang singularity, so as to obey a second law of thermodynamics. As such, it would display precisely the behavior cited above by the Aguirre-Gratton, Bojowald, and Zeh-Kiefer approaches. Physically, the question is whether, in our universe, various arrows of time (thermodynamic, cosmological, electromagnetic, psychological) align. Philosophically, the question is whether there is an underlying metaphysical time which the physical times manifest in various ways.[24] Extendibility, by this understanding, would be a technical artifact rather than an indication of past eternality.

Positing the beginning of classical time at a past boundary may have a dramatic implication first pointed out by philosopher Quentin Smith. Smith's comments were aimed at Hawking's interpretation of the imaginary time featured in the Hartle-Hawking quantum gravity cosmogony. Hawking was attempting to posit an "imaginary" time phase for the universe that transitioned into a later, "normal" time phase. Says Smith:

> ….. such an interpretation is implicitly logically self-contradictory [in featuring a 'time' axis that starts off just like a space dimension, but then transforms to its familiar properties]. The problem appears in the statement that the four-

> dimensional space joins on to the real (Lorentzian) spacetime 'once' (*i.e.*, after) the quantum smearing effects subside: 'The question then arises as to the geometry of the four-dimensional space which has to somehow smoothly join onto the more familiar spacetime once the smearing effects subside.' If the four-dimensional space does not possess a real time value, how can it stand in relation to the four dimensional spacetime of being earlier than it? If the four dimensional space is not in real (Lorentzian) time, then it is not really earlier than, later than, or simultaneous with the four dimensional spacetime manifold. Accordingly, it is false that the 4-sphere joins onto the familiar spacetime once (*i.e.*, after in real time) the quantum effects dissipate [Smith, 1993, p. 318].

Smith later showed [Smith, 1997; 2000] that these two (actually three; there is a transition region) "stages" of the H-H model could coherently co-exist in the same model only if they stood in a more primitive "topological" relation,[25] rather than a temporal one. Thus, if two stages of a cosmogony did not stand on *one and the same* time axis, then it would be false to ascribe to them an *earlier than* or *later than* relation, even though extendibility existed between the relevant universe phases. Recall, for example, that Penrose denies a real time value to his boundary as a physical necessity. When Penrose says, "the remote future of one phase of the universe becomes the Big Bang of the next," his statement is arguably incoherent in the context of his model.

It seems to us Smith's insight applies in general to cosmological models whether they are classical, general relativistic models (which thus include singularities) or quantum gravity models (which include "strong quantum regions"). The "first" phase (which, alternatively, could be a "transition" boundary between different classical or semi-classical spacetimes) by definition does not possess a time value. As Smith points out, because a model phase that lacks a real time value is not before, simultaneous with, or after any other universe state (with respect to a phase that does have real time values) it is not correct to say that the universe evolved from some state "at" the boundary into some state with a real time value.

Very similar language is used by loop quantum theorists themselves.[26] On Smith's view (similarly, Ashtekar *et al.*), it seems correct to view this era as a primitive manifold with topological relations only between the "cycles." The cycles in the "earliest" era seem to be a set of disconnected spacetimes which lack temporal ordering. One could just as well razor[27] these out. One is left with a spacetime with a finite past timeline and no first moment, which grows into a macroscopic universe similar to ours.

Thus it seems 'singularities' are resolved, or not. If so, one has a pre-Big Bang model and faces issues such as a disappearing arrow of time, a reversed arrow of time, and/or entropy problems. If not (meaning time 'dissolves'), then there are no 'pre-Big Bang' phases, since those phases are incommensurate with respect to our FRW-like universe.

One recalls Pitts' insistence on a first moment as a necessary condition of a beginning of time [Pitts, 2008, p. 677]. Precisely the opposite situation may obtain, generically, in

models of the types discussed, despite the technical extendibility of geodesics through the boundary. We suggest that this situation represents a creation *ex nihilo* just as surely as the Big Bang singularity in standard cosmology itself represents an absolute beginning.[28]

In summary, then, it seems to us that a strong case for a beginning of the universe can be made with respect to pre-Big Bang cosmogonies (only some of which involve non-singular models), so that (9) is dispensable so far as premise (2) of the *kalām* cosmological argument is concerned, regardless of how critical it might be to Pitts' imagined singularity argument.

5.0 The beginning of the universe

But what, we may ask, is the nature of the universe's beginning if it is not a GTR-type singularity? The most widely discussed framework for addressing this problem is the Vilenkin "tunneling from nothing" or the Hartle-Hawking "no-boundary" model.[29] Vilenkin observes,

> Many people suspected that in order to understand what actually happened in the beginning, we should treat the universe quantum-mechanically and describe it by a wave function rather than by a classical spacetime. This quantum approach to cosmology was initiated by DeWitt and Misner, and after a somewhat slow start received wide recognition in the last two decades or so. *The picture that has emerged from this line of development is that a small closed universe can spontaneously nucleate out of nothing, where by 'nothing' I mean a state with no classical space and time.* The cosmological wave function can be used to calculate the probability distribution for the initial configurations of the nucleating universes. Once the universe nucleates, it is expected to go through a period of inflation, driven by the energy of a false vacuum. The vacuum energy is eventually thermalized, inflation ends, and from then on the universe follows the standard hot cosmological scenario [Vilenkin, 2002, p. 2; our emphasis].

Hartle and Hawking also claim that their cosmogonic model can be interpreted in such a way that on that interpretation the universe came into being out of "nothing":

> One can interpret the functional integral over all compact four-geometries bounded by a given three-geometry as giving the amplitude for that three-geometry to arise from a zero three geometry; that is, a single point. In other words, the ground state is the probability for the Universe to appear from nothing [Hartle and Hawking, 1983, p. 2961].

On these models the universe clearly has a beginning; but what does it mean to "nucleate out of nothing?" It seems to us that there are four possible ways of interpreting this notion:

1. The "initial" null topology[30] represents literal nothingness. One could imagine that "nothing" as employed by Hartle-Hawking has exactly the relevant meaning for the *kalām* cosmological argument, namely, the positive assertion of nonbeing. If this is what is meant here, then models of this sort feature an unambiguous *creatio ex nihilo*. The second premise of the *kalām* cosmological argument is upheld, and the debate shifts to the first premise.
2. The "initial" null topology is to be construed purely instrumentally. Hawking himself seems to give good grounds for treating Hartle and his proposal as a FAPP (*i.e.*, "for all practical purposes") approach only. The H-H formalism considers a superposition of all possible quantum geometries, with our universe emerging as the most probable, and uses Richard Feynman's "sum over histories" approach to quantum mechanics to predict universe observables. This formalism treats time, in the quantum gravity era, as a fourth spatial dimension. But in his collaboration with Roger Penrose, *The Nature of Space and Time,* Hawking employs the same mathematical approach (analytic continuation) to describe pair production of electron/positron pairs in a strong electric field [Hawking and Penrose, 1996, p. 54]. This is a standard mathematical technique sometimes used when complex analytic functions are better behaved in a certain domain than their real counterparts. It does not imply ontological commitment to the alternative description, however. It seems to us that given the unintelligibility of the "imaginary time" region in these models, it is most reasonable to treat this regime as a useful fiction. Thus, the model serves only to reinforce the second premise of the *kalām* cosmological argument.
3. The "initial" null topology is literally something. Attempts to take the imaginary time interpretation realistically can run into problems similar to those Smith pointed out [Smith 1993, 1997]. In fact, more recent work using computer simulations of the evolution of spacetime suggests that classical spacetime cannot have emerged from a Euclidean 4-space such as is envisioned in the H-H model. Renate Loll, Jan Ambjorn, and Jerzy Jurkiewicz explain:

> In our search for loopholes and loose ends in the Euclidean approach, we finally hit on the crucial idea, the one ingredient absolutely necessary to make the stir fry come out right: the universe must encode what physicists call causality. Causality means that empty spacetime has a structure that allows us to distinguish unambiguously between cause and effect. It is an integral part of the classical theories of special and GTR.
>
> Euclidean quantum gravity does not build in a notion of causality. The term 'Euclidean' indicates that space and time are treated equally. The universes that enter the Euclidean superposition have four spatial directions instead of the usual one of time and three of space. Because Euclidean universes have no distinct notion of time, they have no structure to put events into a specific order; people living in these universes would not have the words 'cause' or 'effect' in their vocabulary. Hawking and others taking this approach have said that 'time is imaginary,' in both a mathematical sense and a colloquial one. Their hope was that causality would emerge as a large-scale property from

> microscopic quantum fluctuations that individually carry no imprint of a causal structure. But the computer simulations dashed that hope [Lolle, Ambjorn, and Jurkiewicz, 2008, p. 3].

In their Causal Dynamical Triangulations approach, causality is demonstrated to be a necessary *fundamental* feature for a cosmogony yielding our observable universe. Lee Smolin explains the implication:

> Some of the most widely believed ideas about quantum gravity are in fact wrong. For example, Stephen Hawking and others used to argue that causal structure was inessential, and that calculations could be done in quantum gravity by ignoring the differences between time and space–differences that exist even in relativity theory–and treating time as if it were another dimension of space [imaginary time] ….. Ambjorn and Lolle's results show that this idea is wrong.
>
> ….. In particular it was shown ….. that if no restriction respecting causality is put in, no classical spacetime geometry emerges …..
>
> One of the rules that Loll and Ambjorn impose is that each quantum spacetime has to be seen as a sequence of possible spaces that succeed one another, like the ticks of a universal clock. The time coordinate, it is argued, is arbitrary, as in GTR, but the fact that the history of the world can be seen as a succession of geometries that succeed one another is not [Smolin 2006, pp. 242–43].

Interestingly, the trajectory of quantum gravity research is tending to correct the "imaginary time" approach and in so doing to affirm the importance of consistent *earlier than / later than* relations and fundamental causality.

4. The "initial" null topology is misconceived. J. Richard Gott and Li Xin Li have criticized Vilenkin (and Hartle-Hawking's) *creatio ex nihilo* approach on two grounds. First, transitions in QM are always between allowed classical states. But Vilenkin and Hartle-Hawking's approach has a transition from a classically forbidden region to a classically allowed region. Second, the Vilenkin and Hartle-Hawking approaches should contain realistic energy fields (something closer to what we actually see in nature). If they did, then Heisenberg's uncertainty principle would require that the initial state of their models have a finite and non-zero energy. If that is the case, then semiclassical quantum models like Vilenkin's and Hartle-Hawking's actually start in a classically allowed, metastable state, rather than "nothing." Gott and Li elaborate:

> The problem with this model [Vilenkin and Hartle-Hawking] is that it ignores the 'zero-point energy.' If there is a conformal scalar field ϕ, then the 'energy' levels should be $E_n=n+1/2$. Even for $n=0$ there is a 'zero-point-energy.' The potential makes the system behave like a harmonic oscillator in the potential well near $a=0$. A harmonic oscillator cannot sit at the bottom of the potential well—the uncertainty principle would not allow it. There must be some zero-point-energy and the particle must have some momentum, as it oscillates within the potential well when the field ϕ is included. Thus, when the 'zero point-energy' is considered, we see that the initial state is not a point but a tiny

> oscillating ($0<a<a_1$) big bang universe, that oscillates between big bangs and big crunches (though the singularities at the big bangs and big crunches might be smeared by quantum effects). This is the initial *classical* state from which the tunneling occurs. *It is metastable, so this oscillating universe could not have existed forever: after a finite half-life, it is likely to decay.* It reaches maximum radius a_1, and then tunnels to a classical de Sitter state at minimum radius a_2 where $a_2<a_0$ [Gott and Li, 1998, p. 38, our emphasis].

Thus, we seem to have the same sort of situation that we encountered with respect to the Emergent model with its associated metastable ESS. The universe cannot be past eternal because the initial metastable state can have had only a finite lifetime. The Gott-Li interpretation seems to be a reasonable option for a realist interpretation of these models.[31] It employs known, meaningful interpretations of physical phenomena from "classical" quantum theory and extends them to the quantum gravity models. One avoids the problems associated with the novelty of asserting a zero-energy condition for the initial state (denied by the Heisenberg uncertainty principle), the novelty of asserting a quantum transition from a forbidden to a classically allowed state (normal quantum theory includes only transitions over or through forbidden regions from one allowed state to another), and it is consistent with more realistic energy fields. This alternative is, of course, also consistent with the second premise of the *kalām* cosmological argument.

6.0 Concluding remarks

In sum, we have seen no good reason to deny that spacetimes featuring an initial cosmological singularity have a beginning; but even more fundamentally we have seen that the claim that the universe began to exist does not depend upon the singular nature of spacetime. Moreover, our survey of contemporary cosmological thinking reveals significant evidence that non-singular models may themselves be incompatible with an infinite past, and therefore singularity-free quantum gravitational approaches can be quite supportive of the *kalām* cosmological argument's second premise. So when Pitts worries, "It is not implausible that some singularity-wielding theistic apologists will be tempted to resist scientific progress in the form of a new quantum theory of gravity in order to maintain an apologetic strategy in which they have invested" [Pitts, 2008, p. 19], one cannot help but wonder who is the object of so uncharitable an allegation. For we have seen that the second premise of the *kalām* cosmological argument enjoys considerable empirical support even if spacetime is non-singular.

Acknowledgments

We thank Martin Bojowald, Thomas Banks, George Ellis, Alexander Vilenkin, Donald Page, Quentin Smith, Timothy McGrew, Rüdiger Vaas and Christian Böhmer for useful discussions and comments.

Notes

1. Writing in the *Cambridge Companion to Atheism*, Quentin Smith reports, "a count of the articles in the philosophy journals shows that more articles have been published about ….. the *Kalām* argument than have been published about any other philosopher's contemporary formulation of an argument for God's existence" [Smith, 2007, p. 183]. For the history of the argument see Craig [1980]; for literature and a recent defense of the argument see Craig and Sinclair [2009].
2. In places Pitts recognizes this: "The truth of the second premise [of the *kalām* cosmological argument], or rather, the source of warrant for the second premise if it is true, is the key question" [Pitts, 2008, p. 687; *cf.* p. 676].
3. So to argue (while avoiding the fallacy of composition) would be to challenge the truth, not of (2), but of (1) of the *kalām* cosmological argument. Sometimes Pitts does conflate what are really challenges to (1) with putative objections to (2). For example, his mention of four-dimensionalism as rendering spacetime "self-explanatory" [Pitts, 2008, p. 688] is a case in point. Again, Pitts' Cosmic Destroyer argument is really an objection to (1): if the coming-into-being of the universe must have a cause, then so must its ceasing-to-be.
4. One cannot, in passing, help but wonder how imposing topological constraints on the doctrine of creation in time, as Pitts would do, is modally enriching, as he claims, since then there are possible worlds (like, apparently, the actual world) in which the universe lacks a first moment and so cannot be said on Pitts' account to have been created in time by God.
5. It is noteworthy that Earman, whose objection Pitts echoes, does not in the end deny the acceptability of our explication of "begins to exist" but merely advises that on such a reading premise (1) "is not an obvious 'metaphysical truth';" in particular (1) does not follow from the principle that *every event has a cause,* which is satisfied in general relativistic cosmology [Earman, 1995, p. 208].
6. *N.B.* that both Pitts's argument and our response are limited to a faux universe where GTR is the fundamental theory of gravity.
7. We also note that there is no good reason to think that metric conventionalism holds in the actual world. As has been argued elsewhere [Craig, 2001, chap. 2], metric conventionalism is an implausible thesis for which no good arguments exist.
8. It is at least worth noting that Pitts, not content with the promise of current quantum cosmology, goes so far as to champion the rights of *unborn* and, indeed, *unconceived* hypotheses as grounds for scepticism about (9):

   > The relevant set of competitors for GTR includes the set of theories that agree with GTR on all experiments to date, whether already entertained on Earth or not. This set might be infinite, might well be large, likely contains several members, and almost certainly has at least one member, a quantum theory of gravity. The set most likely has at least one member that resolves the singularities of GTR. Thus, in the strong field regime it is not at all clear why one should take GTR seriously [Pitts, 2008, p. 696].

   Pitts' final sentence is a *non sequitur*. As Timothy McGrew has remarked in personal conversation, in order to mount a significant challenge to GTR it is not enough to show that it is probable that an unknown theory of quantum gravity exists which

shares GTR's empirical adequacy to date and resolves the singularities of GTR. Pitts also needs to show that this unknown competing theory has, by comparison to GTR, a non-negligible prior probability. For there may be other overriding factors that are relevant to its probability, for example, its want of simplicity or elegance. Since these theories are unconceived, it is hard to see how Pitts can know such a thing. In any case, as we remark in the text, the important issue is not whether that unknown theory resolves the singularities of GTR but whether the physical universe as described by the theory is infinitely extrapolable to the past. What one wants from Pitts is some evidence or argument that there is an unknown theory of quantum gravity meeting all the *desiderata*.

9. Our intention is to provide a complementary update to the discussion in [Craig & Sinclair 2009]. Thus, for example, the material in the original essay had significant content regarding string based cosmogonies which will not be repeated here.
10. It also permits us to regard records of the past, such as our own memories, as being relics of a true reality, as opposed to illusory. We will hold that starlight that appears in our telescopes in the present actually came from a distant star in a real past. So called 'top-down' interpretations of the Feynman path integral approach to quantum gravity [Hawking, 2003], or 'decoherent histories' (see [Hartle 1998] & [Dowker & Kent, 1996], are examples of quantum gravity approaches that would deny the objective reality of a unique past that *evolved* into the present. This doesn't mean, necessarily that these approaches must be rejected. Rather, we reject the *interpretation* of the formalism. Hawking, in the above reference, says about his own theory that "One can interpret this in the bottom up picture as the spontaneous creation of an inflating universe from nothing." On this view, we have precisely the creation *ex nihilo* that lends evidence towards *kalām's* second premise.
11. This implication was affirmed by Vilenkin in a personal communication on March 4, 2004. Note that this conclusion would apply to the well-known Ekpyrotic cosmogony of Paul Steinhardt and Neil Turok mentioned by Monton. Their model has a genuine past boundary and thus accords with the second premise of the *kalām* cosmological argument. See [Craig and Sinclair, 2009] for a more detailed explanation.
12. See also the collection of papers by Brett McInnis [McInnis, 2007a], [McInnis, 2007b], [McInnis, 2008]. McInnis provides an alternative argument for why inflation must have an ultimate origin. He suggests that in string theory (for purely geometric reasons), the arrow of time can only originate in a topological creation event 'from nothing' such as in the Vilenkin proposal (see section 5 of our paper). Subsequent bubble universes, whether embedded or distinct disconnected spacetimes (see [Carroll and Chen 2004]), will have an arrow of time only if they *inherit* it from their mother. Thus, given our observed arrow of time, we know we have an ultimate 'mother' that came into being via creation *ex nihilo*.
13. To illustrate, one of the authors once participated in a similar exercise while working in the Defense industry. The author noted that a particular air-to-air combat model explicitly included missile flyouts against air targets. A much more elegant (and efficient) model was built by deriving a set of equations whereby time and distance intervals could be dropped from the formalism. The combat could be described by an impact sequence of missiles salvos that does not reference the time or space intervals between salvo arrivals. Now would the mere fact of the existence of this model (and its superiority in terms of simplicity, explanatory scope, *etc*.) imply that time and space intervals do not exist? This would be a ludicrous conclusion, as any experience

as a combat aviator would demonstrate. Air superiority is maintained by having the superior range missile (a distance interval), or shortest flight (time interval) such that one's missile gets there *first*. It is precisely the intervals that produce salvo impact order and hence air superiority. Individual life and death, even the rise and fall of empires, ride on this conclusion. One is reminded of Einstein's advice that we should make our models as simple as possible, but not *too* simple.

14. See [Aguirre 2007] for a detailed attempt to circumvent BGV in favor of a time-reversed or 'emergent' non-singular cosmogony. These suggestions of Aguirre's are considered in section 4.4 & 4.6.
15. See, *e.g.*, [Earman and Mosterin, 1999] for a related argument.
16. See also [Damour and Henneaux, 2000] whose results update B-K-L to consider quantum string cosmologies: " ….. our findings suggest that the spatial inhomogeneity continuously increases toward a singularity, as all quasi-uniform patches of space get broken up into smaller and smaller ones by the chaotic oscillatory evolution. In other words, the spacetime structure tends to develop a kind of 'turbulence.'"
17. Ellis does indicate that a finite past is an acceptable possible form for an Emergent model [Ellis and Maartens, 2004, section V, p. 5].
18. The quantum state must not be infinitely squeezed to be semiclassical (see section 4.5); the center equilibrium point only exists if the semiclassical assumption holds.
19. See [Craig & Sinclair 2009] for an extensive discussion of another asymptotically static cosmogony, the string based pre-Big Bang inflation model of Maurizio Gasperini & Gabrielle Veneziano. See [Kaloper et al. 1998] for a similar criticism based on fluctuation of the dilaton field.
20. Also see [Tavakol and Carneiro, 2009]. The argument is that the past is infinite. But a suggested method of leaving the stabilized ESS so as to allow an inflationary expansion invokes a quantum tunneling from a local minimum in potential (stabilized ESS) to a global minimum (open ended inflation). But this local minimum is a metastable state with a finite lifetime. Thus the model is not past infinite.
21. Examples in the recent literature of asymptotically static non-singular cosmogonies include, *e.g.*, [Falciano and Pinto-Neto, 2009], [Seahra and Böhmer, 2009], and [Barrow and Tsagas, 2009]. The first compares the predictions of such models with quantum scalar perturbations observed in the cosmic background radiation via the Wilkinson Microwave Anisotropy Probe and casts doubt on their ultimate viability. The second and third investigate the viability of a past eternal ESS model. The Seahra and Böhmer paper shows that an ESS is unstable to some type of perturbation for generic quantum gravity theories and generic equation of state (that is, the type of matter/energy residing in the universe) for f(R) gravity theories. The Barrow and Tsagas paper exploits a stability region for ESS for "phantom" or "ghost" matter universes. See [Craig and Sinclair, 2009] for a discussion of some of these "phantom bounce" cosmogonies.
22. Banks complains,

> "I have a problem with ALL cyclic cosmologies ….. The collapsing phase of these models always have a time dependent Hamiltonian for the quantum field fluctuations around the classical background. Furthermore the classical backgrounds are becoming singular. This means that the field theories will be excited to higher and higher energy states (define energy in some adiabatic fashion during the era when the cosmology is still fairly slowly varying, and use this to classify the states, even though it is not conserved). High energy states in field theory have the ergodic

property—they thermalize rapidly, in the sense that the system explores all of its states. Willy Fischler and I proposed that in this situation you would again tend to maximize the entropy. We called this a Black Crunch and suggested the equation of state of matter would again tend toward $p = \rho$. It seems silly to imagine that, even if this is followed by a re-expansion, that one would start that expansion with a low entropy initial state, or that one had any control over the initial state at all" (pers. comm., Oct 12, 2008).

An important caveat is that a general description of entropy in the presence of gravity (either in GTR or the still undiscovered Quantum Gravity) has never been achieved. But there are some special cases that are thought reliable by most cosmologists [see Carroll, 2007, slides 11 and 12], such as the entropy of a thermal gas, black holes, and cosmic horizons (this last being related to a positive cosmological constant). These three cases are thought to represent the dominant entropy contributors to the universe for three different eras: the early universe (thermal gas), the present (black holes), and the far future (a "pure" de Sitter-like cosmic horizon). If these special cases are accurate, they demonstrate that entropy increases as the universe ages, thus providing some proof of a generalized second law of thermodynamics.

23. See also [Coule 2008], where a suggestion is made that maximum entropy is capped by the holographic bound (cosmic horizon entropy for a de Sitter universe), and that the entropy of a bounce violates this restriction.
24. A philosopher sympathetic with a tensed theory of time would have an additional rationale for believing that such models really represent a multiverse propagating from a common origin. There would be a reason why the cosmological, thermodynamic, electromagnetic, and psychological arrows of time would all align; they are physical manifestations of the same underlying temporal becoming of metaphysical time.
25. Note that Smith and Pitts are interested in different topological features of spacetime. Pitts's concern is whether the universe has a first moment. Smith's concern is whether the topology of a spacetime manifold (*e.g.*, unbroken sheet, torus, figure eight, *etc.*) is more fundamental than its metrical aspects.
26. "….. the question of whether the universe had a beginning at a finite time is now 'transcended'. At first, the answer seems to be 'no' in the sense that the quantum evolution does not stop at the big bang. However, since space-time geometry 'dissolves' near the big-bang, *there is no longer a notion of time, or of 'before' or 'after' in the familiar sense*. Therefore, strictly, the question is no longer meaningful. The paradigm has changed and meaningful questions must now be phrased differently, without using notions tied to classical space-times" [Ashtekar, Bojowald and Lewandowski, 2003, pg. 263, our emphasis].
27. "And why would we not apply Occam's razor to the pre-history?" (Bojowald 2009, p. 16)
28. A creation *ex nihilo* makes more sense than positing time's emergence "from" a pure quantum state. Augustine seems to have had a similar insight 1600 years ago:

> "We can correctly say, 'There was a time when Rome did not exist: there was a time when Jerusalem, or Abraham, or man, or anything of this kind did not exist'. We can in fact say, 'There was a time when the world did not exist', if it is true that the world was created not at the beginning of time, but some time after. But to say, 'There was a time when time did not exist', is as nonsensical as to say, 'There was a man when no

man existed', or, 'This world existed when the world was not'" [Augustine, 1984, *City of God*, Bk XII, ch 16].

29. Also see [Mcinnis 2007a, 2007b, 2008] for a string based model of this type.
30. "Creation of a universe from nothing ….. is a transition from the null topological sector containing no universes at all to the sector with one universe of topology S^3" [Vilenkin 1994, p. 23].
31. In [Craig and Sinclair, 2009], we profile Gott and Li's own attempt at solving cosmology's origins problem based on the postulate of closed timelike curves (a time machine). Unfortunately, they can avoid Stephen Hawking's Chronology Protection Conjecture only by asserting an "initial" cosmic vacuum which is measure zero with respect to all possible configurations. In other words, their model is infinitely fine-tuned.

References

Aguirre, Anthony and Steve Gratton. (2002): "Steady State Eternal Inflation," *Physical Review D65 (2002) 083507*, preprint: http://arxiv.org/abs/astro-ph/0111191v2.

Aguirre, Anthony (2007): "Eternal Inflation Past and Future", in *Beyond the Big Bang* Ed. R. Vaas, Heidelberg: Springer Verlag, preprint: http://arxiv.org/abs/0712.0571

Al-Ghazali (1962). *Kitab al-Iqtisad fi'l-Iqtiqad.* Ankara: University of Ankara Press.

Al-Ghazali (1963). *Tahafut al-Falasifah*. Trans. S. A. Kamali. Lahore: Pakistan Philosophical Congress.

Ashtekar, A, Bojowald, M and and Lewandowski, J (2003): "Mathematical structure of loop quantum cosmology", in *Adv.Theor.Math.Phys*. 7 233–268, preprint: http://arxiv.org/abs/gr-qc/0304074

Ashtekar, Abhay, Pawlowski, Tomasz, and Singh, Parampreet (2006), "Quantum Nature of the Big Bang", in Phys. Rev. D 74, 084003, Preprint: http://arxiv.org/abs/gr-qc/0602086v2

Asimov, Isaac (1989): "The Relativity of Wrong", in *The Skeptical Inquirer*, volume 14 No. 1, Fall 1989, available on the web at: http://chem.tufts.edu/AnswersInScience/RelativityofWrong.htm

Augustine. (1984) *City of God*, London, Penguin Books.

Banks, T. and Fischler, W. (2002): "Black Crunch," preprint: http://arxiv.org/abs/hep-th/0212113v1

Barrow, J. and Dabrowski, M. (1995): "Oscillating universes," *Monthly Notices of the Royal Astronomical Society* 275, 850–62.

Barrow, J. and Tsagas, C. (2009) "On the Stability of Static Ghost Cosmologies," accessed on the web at: http://arxiv.org/abs/0904.1340v1 (May 21, 2009).

Belinsky, V. A., Khalatnikov, I. M., and Lifshitz, E. M. (1970): "Oscillatory approach to a singular point in the relativistic cosmology," *Advances in Physics* 19, 525–73.

Bojowald, M., Maartens, R. and Singh, P. (2004): "Loop Quantum Gravity and the Cyclic Universe", *Phys.Rev. D*70 (2004) 083517, preprint: http://arxiv.org/abs/hep-th/0407115

Bojowald, M., Date, G., and Hossain, G. (2004), "The Bianchi IX model in Loop Quantum Cosmology", in *Class. Quant. Grav. 21, 3541*, Preprint: http://arxiv.org/abs/gr-qc/0404039

Bojowald, M. (2005): "Original questions," *Nature* 436, 920–1.

Bojowald, M. (2006): "Universe scenarios from loop quantum gravity", in *Annalen Phys*. 15: 326–341, preprint: http://arxiv.org/abs/astro-ph/0511557v1

Bojowald, M. (2007): "Large scale effective theory for cosmological bounces", in *Phys Rev D* 74, 081301, preprint: http://arxiv.org/abs/gr-qc/0608100v2

Bojowald, M. and Tavakol, R. (2008a): "Recollapsing quantum cosmologies and the question of entropy", in *Phys. Rev. D* 78, 023515, Preprint: http://arxiv.org/abs/0803.4484v1

Bojowald, M. and Tavakol, R. (2008b): "Loop Quantum Cosmology: Effective theories and oscillating universes" In *Beyond the Big Bang.* Ed. R. Vaas, Heidelberg: Springer Verlag, Preprint: http://arxiv.org/abs/0802.4274v1

Bojowald, M. (2009): "A Momentous Arrow of Time", in *The Arrow of Time*, Ed. L. Mersini-Houghton, Heidelberg: Springer Verlag.

Borde, A., Guth, A., and Vilenkin, A. (2003): "Inflationary spacetimes are not past-complete," *Physical Review Letters* 90, 151301, preprint: http://arxiv.org/abs/gr-qc/0110012.

Carroll, Sean (2008): "What if Time Really Exists?", FXQI essay contest on the nature of time, available on the web at: http://www.fqxi.org/community/forum/topic/318

Carroll, Sean *et al* (2007) "Why is the Past Different From the Future: The Origin of the Universe and the Arrow of Time", available on the web at: http://preposterousuniverse.com/talks/time-colloq-07/

Carrol, Sean and Chen, Jennifer (2004): "Spontaneous Inflation and the Origin of the Arrow of Time", preprint: http://arxiv.org/abs/hep-th/0410270v1

Coleman, S. and De Luccia, F. (1980): "Gravitational effects on and of vacuum decay," *Physical Review* D21, 3305.

Copan, P. and Craig, Wm. L. (2004): *Creation out of Nothing.* Grand Rapids, Mich.: Baker.

Coule, D.H. (2008), "Holography constrains quantum bounce", preprint: http://arxiv.org/abs/0802.1867

Craig, Wm. L. (2006): "Naturalism and Cosmology." In *Analytic Philosophy without Naturalism*. Ed. A. Corradini, S. Galvan, and E. J. Lowe, pp. 97–133. Routledge Studies in Contemporary Philosophy. New York: Routledge.

Craig, Wm. L. (1980). *The Cosmological Argument from Plato to Leibniz*. London: Macmillan & Co.

Craig, Wm. L. (2001): *Time and the Metaphysics of Relativity*. Philosophical Studies Series 84. Dordrecht: Kluwer Academic Publishers.

Craig, Wm. L. and Sinclair, Jas. (2009): "The *Kalām* Cosmological Argument." In *Blackwell Companion to Natural Theology*, pp. 101–201. Ed. Wm. L. Craig and J. P. Moreland. Oxford: Blackwell.

Damour, T. and Henneaux, M. (2000): "Chaos in superstring cosmology," *Physical Review Letters* 85, 920–3, preprint: http://aps.arxiv.org/abs/hep-th/0003139.

Dowker, F. and Kent, A. (1996): "On the Consistent Histories Approach to Quantum Mechanics", *J.Statist.Phys.* 82 (1996) 1575–1646, preprint: http://arxiv.org/abs/gr-qc/9412067v2

Earman, J. (1995): *Bangs, Crunches, Whimpers, and Shrieks: Singularities and Acausalities in Relativistic Spacetimes*. New York: Oxford University Press.

Earman, J. and Mosterin, J. (1999): "A critical look at inflationary cosmology," *Philosophy of Science* 66, 1–49.

Ellis, G. F. R., Murugan, Jeff and Tsagas, Christos (2004): "The Emergent Universe: An Explicit Construction", In *Class.Quant.Grav.21*:233–250, preprint: http://arxiv.org/abs/gr-qc/0307112

Ellis, G. F. R. and Maartens, R. (2004) "The emergent universe: inflationary cosmology with no singularity," *Classical and Quantum Gravity* 21, 223, preprint: http://arxiv.org/abs/gr-qc/0211082.

Falciano, F. T. and Pinto-Neto, N. (2009): "Scalar Perturbations in Scalar Field Quantum Cosmology," *Physical Review D*79: 023507, 2009, preprint: http://arxiv.org/abs/0810.3542

Geroch, R. (1978): *GR From A to B*, Chicago & London, University of Chicago Press.

Gott, J. R. III and Li, L-X. (1998): "Can the universe create itself?" *Physical Review D* 58: 2, 023501–1.

Hartle, J. (1998): "Quantum Pasts and the Utility of History", *Phys.Scripta* T76 (1998) 67, preprint: http://arxiv.org/abs/gr-qc/9712001v1

Hawking, S. (2003): "Cosmology from the Top Down": Talk presented at the Davis Inflation Meeting, 2003, preprint: http://arxiv.org/abs/astro-ph/0305562

Hawking, S. and Hartle, J. (1983) "The wave function of the universe," *Physical Review D* 28: 12, 2960–75.

Hawking, S. and Penrose, R. (1970) "The singularities of gravitational collapse and cosmology," *Proceedings of the Royal Society of London A* 314, 529–48.

Hawking, S. and Penrose, R. (1996) *The Nature of Space and Time*. Princeton, N. J. : Princeton University Press.

Jurkiewicz, J., Loll, L. & Ambjorn, J. (2008): "Using Causality to Solve the Puzzle of Quantum Spacetime," *Scientific American*, July, found on the web at: http://www.scientificamerican.com/article.cfm?id=the-self-organizing-quantum-universe.

Kaloper, N., Linde, A. & Bousso, R. (1998) "Pre-Big-Bang Requires the Universe to be Exponentially Large From the Very Beginning," *Physical Review D, Volume 59, 043508*, accessed online at hep-th/9801073.

Kiefer, C. and Zeh, H. D. (1995): "Arrow of time in a recollapsing quantum universe", *Phys.Rev. D51* 4145–4153, preprint: http://arxiv.org/abs/gr-qc/9402036v2

Kiefer, Claus (2008): "Does Time Exist in Quantum Gravity?", FXQI essay contest on the Nature of Time, available on the web: www.fqxi.org/data/essay-contest-files/Kiefer_fqx.pdf

Kiefer, Claus (2009): "Can the Arrow of Time Be Understood from Quantum Cosmology?", in *The Arrow of Time*, Ed. L. Mersini-Houghton, Heidelberg: Springer Verlag, preprint: http://arxiv.org/abs/0910.5836v1

Loll, R. (2007): "The Emergence of Spacetime, or, Quantum Gravity on Your Desktop", in *Classical Quantum Gravity* Volume 25, number 11 114006, available on the web at: http://arxiv.org/abs/0711.0273v2

Mersini-Houghton, Laura (2009): "Notes on Time's Enigma", FXQI conference on time, Azores, 2009, available on the web at: http://arxiv.org/abs/0909.2330v1

McInnis, Brett (2007a): "Arrow of Time in String Theory", In *Nucl.Phys*. B782:1–25, preprint: http://arxiv.org/abs/hep-th/0611088v3

McInnis, Brett (2007b): "The Arrow of Time in The Landscape", To appear in R. Vaas (ed.): *Beyond the Big Bang*. Springer: Heidelberg 2010, preprint: http://arxiv.org/abs/0711.1656v2

McInnis, Brett (2008): "Initial Conditions for Bubble Universes", In *Phys.Rev.D77*:123530, preprint: http://arxiv.org/abs/0705.4141v5

Monton, Bradley (forthcoming). "Prolegomena to Any Future Physics-Based Metaphysics." *Oxford Studies in Philosophy of Religion* 3.

Mulryne, David, Reza Tavakol, James E. Lidsey and George F. R. Ellis (2005): "An Emergent Universe From a Loop," *Physical Review D* 71, 123512 (2005), preprint: http://arxiv.org/abs/astro-ph/0502589v1

Oppy, G. (2006): *Arguing about Gods.* Cambridge: Cambridge University Press.

Overbye, D. (2006): "9 billion-year-old 'dark energy' reported." *The New York Times*, January 15, 2008.

Parikh, M., Ivo Savonije and Erik Verlinde. (2003): "Elliptic de Sitter Space," *Physical Review D67* (2003) 064005, preprint: http://arxiv.org/abs/hep-th/0209120v2.

Penrose, Roger (2005): *The Road to Reality*, New York: Alfred A. Knopf.

Penrose, R. (2006): "Before The Big Bang: An Outrageous New Perspective And Its Implications For Particle Physics," *Proceedings of the European Particle Accelerator Conference (EPAC) 2006.* Edinburgh, Scotland, pp. 2759–62, found on the web at: http://accelconf.web.cern.ch/AccelConf/e06/PAPERS/THESPA01.PDF.

Penrose, R. (2009): "Black holes, quantum theory and cosmology", in *J. Phys.: Conf. Ser.* 174 01 2001, available on the web at: http://iopscience.iop.org/1742–6596/174/1/012001/pdf/1742–6596_174_1_012001.pdf

Pitts, J. B. (2008): "Why the Big Bang Singularity Does Not Help the *Kalām* Cosmological Argument for Theism." *British Journal for the Philosophy of Science*: 59, 675–708.

Quine, W. V. (1969): "Epistemology Naturalized." In *Ontological Relativity and Other Essays*, pp. 69–90. New York: Columbia University Press.

Rea, Michael C. (2002): *World without Design: The Ontological Consequences of Naturalism.* Oxford: Clarendon Press.

Seahra, Sanjeev S. and Böhmer, C. G. (2009): "Einstein static universes are unstable in generic f(R) models", *Physical Review D* 79: 064009,2009, preprint: http://arxiv.org/abs/0901.0892

Smith, Q. (1985): "On the Beginning of Time." *Noûs* 19, pp. 579–84.

Smith, Q. (1993): "The Wave Function of a Godless Universe." In *Theism, Atheism and Big Bang Cosmology*, p. 301–37. Ed. Quentin Smith and William Lane Craig. Oxford: Clarendon Press

Smith, Q. (1997): "The Ontological Interpretation of the Wave Function of the Universe." *The Monist*, vol. 80, no.1, pp, 160–185.

Smith, Q. (2000): "The Black Hole Origin Theory of the Universe: Frontiers of Speculative, Current Physical Cosmology". Found on the web at: http://www.faculty.umb.edu/gary_zabel/Courses/Parallel%20Universes/Texts/the_black_hole_origin_theory_of_the_universe_frontiers_of_s.htm

Smith, Q. (2007): "Kalam Cosmological Arguments for Atheism." In *The Cambridge Companion to Atheism*, pp. 182–98. Ed. M. Martin. Cambridge Companions to Philosophy. Cambridge University Press.

Sorabji, R. (1983): *Time, Creation, and the Continuum*. Ithaca, N.Y.: Cornell University Press.

Sorkin, R. D. & Rideout, D.P. (1999): "Classical sequential growth dynamics for causal sets", in *Physical Review D*, Volume 61, 024002, available on the web at: http://arxiv.org/abs/gr-qc/9904062v3

Smolin, L. (2006): *The Trouble With Physics*. New York, N.Y.: Houghton Mifflin.

Smolin, Lee (2009): "The unique universe", in *PhysicsWorld*, June 2009, available on the web at: http://physicsworld.com/cws/article/print/39306

Tavakol, R. and Carneiro S. (2009): "Stability of the Einstein static universe in the presence of vacuum energy", In *Physical Review D* 80, 043528, preprint: http://arxiv.org/abs/0907.4795v2

't Hooft, Gerard (2009): "Quantum Gravity without Space-time Singularities or Horizons", Presented at the Eric Summerschool of Subnuclear Physics, available on the web at: http://arxiv.org/abs/0909.3426

Tod, K. P. (2003): *Class. Quantum Gravity* 20, 521–534.

Tolman, R. C. (1934): *Relativity, Thermodynamics and Cosmology*. Oxford: Clarendon Press.

Vaas, R. (2004): "Time before time: classifications of universes in contemporary cosmology, and how to avoid the antinomy of the beginning and eternity of the world." In W. Löffler and P. Weingartner (eds.), *Knowledge and Belief. Papers of the 26th International Wittgenstein Symposium*, 351–3. Kirchberg am Wechsel: Austrian Ludwig Wittgenstein Society, preprint: http://arXiv.org/abs/physics/0408111.

Veneziano G. and Gasperini, M. (2002): "The pre big bang scenario in string cosmology," *Physics Reports* 373, 1, preprint: http://arxiv.org/abs/hep-th/0207130.

Vilenkin, A. (1982): "Creation of Universes from Nothing", *Phys. Lett*. 117B, 25–28

Vilenkin, A. (1994): "Approaches to quantum cosmology," *Physical Review D* 50, 2581–94, preprint: http://lanl.arxiv.org/abs/gr-qc/9403010v1.

Vilenkin, A. (2002): "Quantum cosmology and eternal inflation." In *The Future of Theoretical Physics and Cosmology*, Proceedings of the Conference in Honor of Stephen Hawking's 60th birthday, preprint: http://arxiv.org/abs/gr-qc/0204061

Vilenkin, A. (2006): *Many Worlds in One*. New York: Hill & Wang.

Zeh, H. D. (2009): "Open questions regarding the arrow of time", in *The Arrow of Time*, Ed. L. Mersini-Houghton, Heidelberg: Springer Verlag, preprint http://arxiv.org/abs/0908.

4

The Beginning of the Universe

Alexander Vilenkin

We live in the aftermath of a great explosion—the big bang—that occurred 13.7 billion years ago. At the time of the big bang, the universe was filled with a fireball, a dense mixture of energetic particles and radiation. For nearly a century, physicists have been studying how the fireball expanded and cooled, how particles combined to form atoms, and how galaxies and stars were gradually pulled together by gravity. This story is now understood in great quantitative detail and is supported by abundant observational data.[1]

The question, however, remains whether the big bang was truly the beginning of the universe. A beginning in what? Caused by what? Caused by what? And determined by what, or whom? These questions have prompted physicists to make every attempt to avoid a cosmic beginning.

In this essay, I review where we now stand.

The Penrose singularity theorem

The problem has been staring us in the face since the early days of scientific cosmology. In the 1920s, the Russian mathematician Alexander Friedmann provided a mathematical description of an expanding universe by solving Albert Einstein's equations of general relativity. Friedmann assumed for the sake of simplicity that the distribution of matter in the universe was perfectly uniform. His solutions had a mystifying feature: as the evolution of the universe is followed backward in time, the density of matter and the curvature of space-time grow without bounds, becoming infinite a finite time ago.[2] The moment of infinite density is a cosmological singularity. At that point, the mathematical expressions appearing in the equations of general relativity become ill defined, and evolution cannot be continued. This would seem to suggest that the universe did have a beginning, but not one describable by the laws of physics.

Physicists hoped initially that the singularity might be an artifact of Friedmann's simplifying assumption of perfect uniformity, and that it would disappear in more realistic solutions of Einstein's equations. Roger Penrose closed this loophole in the mid-1960s by showing that, under a very general assumption, the singularity was unavoidable.[3] Under the null convergence condition, gravity always forces light rays to

converge.[4] This implies that the density of matter or energy measured by any observer cannot be negative. The conclusion holds for all familiar forms of classical matter.

Penrose's proof hinges on the concept of an incomplete geodesic. In general relativity, the trajectories of matter are represented by straight lines in space-time, or geodesics. If space-time is free of singularities, all geodesics should have infinite extent. A geodesic encountering a singularity cannot be further extended. Such geodesics are incomplete. Penrose showed that a space-time satisfying the null convergence condition (and some additional mild assumptions) must contain incomplete geodesics. A singularity appears to be unavoidable.

Penrose's argument was not entirely convincing. Even though the null convergence condition is satisfied by classical matter, quantum fluctuations can create regions of negative energy density.[5] In the extreme conditions near the big bang, quantum fluctuations are of great importance. Penrose's argument no longer applies.

This is where things stood until the early 1980s, when Alan Guth introduced the idea of cosmic inflation.[6]

Eternal inflation

Inflation is a period of super-fast, accelerated expansion in the early history of the universe. In a fraction of a second, a tiny subatomic region blows up to dimensions greater than the entire currently observable universe. The expansion is driven by a false vacuum.

A vacuum is ordinarily thought of as empty space, but according to modern particle physics what is empty is not nothing. The vacuum is a physical object, endowed with energy density and pressure. It can be in a number of different states, or vacua. The properties and types of elementary particles differ from one vacuum to another.

The gravitational force induced by a false vacuum unusual in that it is repulsive. The higher the energy of the vacuum, the stronger the repulsion. This kind of vacuum is unstable. It decays into a low-energy vacuum and the excess energy produces a fireball of particles and radiation. False vacua were not invented for the purposes of inflation. Their existence follows from particle physics and general relativity.

The theory of inflation assumes that at some early time in its history the universe occupied a high-energy false vacuum. Repulsive gravitational forces then caused a super-fast, exponential expansion of the universe. There is a characteristic time in which the size of the universe doubles. Depending on the model, the doubling times can be as short as 10^{-37} seconds. In about 330 doubling times, the universe grows by a factor of 10^{100}. No matter its initial size, the universe very quickly becomes huge. Because the false vacuum is unstable, it eventually decays, producing a fireball, marking the end of inflation. The fireball continues to expand by inertia and evolves along the lines of standard big bang cosmology.

Inflation explained some otherwise puzzling features of the universe, matters that big bang cosmology was forced to assume. It explained the expansion of the universe, its high temperature, and its observed homogeneity. Inflationary theory predicted that

Euclidean geometry describes the universe on the largest scales. It also predicted a nearly scale-independent spectrum of small-density perturbations caused by quantum fluctuations during inflation. These predictions have been confirmed.

The theory of inflation has led to a revision in our view of the universe. Inflation does not come to an end everywhere at once. Regions where the false vacuum decays somewhat later are rewarded by a larger inflationary expansion, so false vacuum regions tend to multiply faster than they decay. In our cosmic neighborhood, inflation ended 13.7 billion years ago; in remote parts of the universe it is still continuing. Regions like ours are constantly being formed. This never-ending process is called eternal inflation. Eternal inflation is generic; and predicted by most models.

False vacuum decay is model-dependent. In this essay, I focus on models in which it occurs through bubble nucleation. Vacuum decay is similar to the boiling of water. Low-energy regions appear as microscopic bubbles and immediately start to grow at a rate rapidly approaching the speed of light. The bubbles are then driven apart by inflationary expansion, making room for more bubbles. We live in one of these bubbles, but we can observe only a small part of it. No matter how fast we travel, we cannot catch up to the expanding boundary of our universe.

Ours is a self-contained universe.

Eternal inflation raises an intriguing possibility. If inflation goes on and on into the future, could it also have gone on and on into the past?[7] A universe without a beginning would make it unnecessary to ask how it began.

As is so often the case in physics, an irresistible force is now about to encounter an immovable obstruction.

The Borde-Guth-Vilenkin theorem

The obstruction may be found in the Borde-Guth-Vilenkin (BGV) theorem.[8] Loosely speaking, our theorem states that if the universe is, on average, expanding, then its history cannot be indefinitely continued into the past. More precisely, if the average expansion rate is positive along a given world line, or geodesic, then this geodesic must terminate after a finite amount of time. Different geodesics, different times. The important point is that the past history of the universe cannot be complete. An outline of the proof is provided in the Appendix.

The BGV theorem allows for some periods of contraction, but on average expansion wins. The volume of the universe increases with time. Inflation cannot be eternal and must have some sort of a beginning.

The BGV theorem is sweeping in its generality. It makes no assumptions about gravity or matter. Gravity may be attractive or repulsive, light rays may converge or diverge, and even general relativity may decline into desuetude: the theorem would still hold.

A number of physicists have constructed models of an eternal universe in which the BGV theorem is no longer pertinent. George Ellis and his collaborators have suggested that a finite, closed universe, in which space closes upon itself like the surface of a

sphere, could have existed forever in a static state and then burst into inflationary expansion.[9] Averaged over infinite time, the expansion rate would then be zero, and the BGV theorem would not apply. Ellis constructed a classical model of a stable closed universe and provided a mechanism triggering the onset of expansion. Ellis made no claim that his model was realistic; it was intended as a proof of concept, showing that an eternal universe is possible. Not so. A static universe is unstable with respect to quantum collapse.[10] It may be stable by the laws of classical physics, but in quantum physics a static universe might make a sudden transition to a state of vanishing size and infinite density. No matter how small the probability of collapse, the universe could not have existed for an infinite amount of time before the onset of inflation.

There is another way that the universe might be eternal in the past. It could have cycled through an infinite succession of expansions and contractions. This notion was briefly popular in the 1930s, but was then abandoned because of its apparent conflict with the second law of thermodynamics. The second law requires that entropy should increase in each cycle of cosmic evolution. If the universe had already completed an infinite number of cycles, it would have reached a state of thermal equilibrium, and so a state of maximum entropy. All the energy of ordered motion would have turned into heat, a uniform temperature prevailing throughout.

We do not find ourselves in such a state.

The idea of a cyclic universe was recently revived by Paul Steinhardt and Neil Turok.[11] They suggested that in each cycle expansion is greater than contraction, so that the volume of the universe is increased. The entropy of the universe we can now observe could be the same as the entropy of some similar region in an earlier cycle; nonetheless, the total entropy of the universe would have increased because the volume of the universe is now greater than it was before. As time goes on, both the entropy and the total volume grow without bounds, and the state of maximum entropy is never reached. There is no maximum entropy.[12]

The problem with this scenario is that, on average, the volume of the universe still grows, and thus the BGV theorem can be applied. This leads immediately to the conclusion that a cyclic universe cannot be past-eternal.

God's proof

Theologians have welcomed any evidence for the beginning of the universe as evidence for the existence of God. "As to the first cause of the universe," wrote the British astrophysicist Edward Milne, "this is left for the reader to insert, but our picture is incomplete without Him."[13] Some scientists feared that a cosmic beginning could not be described in scientific terms. "To deny the infinite duration of time," asserted the Walter Nernst, "would be to betray the very foundations of science."[14]

Richard Dawkins, Lawrence Krauss, and Victor Stenger have argued that modern science leaves no room for the existence of God. A series of science–religion debates has been staged, with atheists like Dawkins, Daniel Dennett, and Krauss debating theists like William Lane Craig.[15] Both sides have appealed to the BGV theorem, both sides appealing to me—of all people!—for a better understanding.

The cosmological argument for the existence of God consists of two parts. The first is straightforward:

- everything that begins to exist has a cause;
- the universe began to exist;
- therefore, the universe has a cause.[16]

The second part affirms that the cause must be God.

I would now like to take issue with the first part of the argument. Modern physics can describe the emergence of the universe as a physical process that does not require a cause.

Nothing can be created from nothing, says Lucretius, if only because the conservation of energy makes it impossible to create nothing from nothing. For any isolated system, energy is proportional to mass and must be positive. Any initial state, prior to the creation of the system, must have the same energy as the state after its creation.

There is a loophole in this reasoning. The energy of the gravitational field is negative;[17] it is conceivable that this negative energy could compensate for the positive energy of matter, making the total energy of the cosmos equal to zero. In fact, this is precisely what happens in a closed universe, in which the space closes on itself, like the surface of a sphere. It follows from the laws of general relativity that the total energy of such a universe is necessarily equal to zero. Another conserved quantity is the electric charge, and once again it turns out that the total charge must vanish in a closed universe.

I will illustrate these statements for the case of an electric charge, using a two-dimensional analogy. Imagine a two-dimensional closed universe, which we can picture as a surface of a globe. Suppose we place a positive charge at the north pole of this universe. Then the lines of the electrical field emanating from the charge will wrap around the sphere and converge at the south pole. This means that a negative charge of equal magnitude should be present there. Thus, we cannot add a positive charge to a closed universe without adding an equal negative charge at the same time. The total charge of a closed universe must therefore be equal to zero.

If all the conserved numbers of a closed universe are equal to zero, then there is nothing to prevent such a universe from being spontaneously created out of nothing. And according to quantum mechanics, any process which is not strictly forbidden by the conservation laws will happen with some probability.[18]

A newly-born universe can have a variety of different shapes and sizes and can be filled with different kinds of matter. As is usual in quantum theory, we cannot tell which of these possibilities is actually realized, but we can calculate their probabilities. This suggests that there could be a multitude of other universes.

Quantum creation is similar to quantum tunneling through energy barriers in quantum mechanics. An elegant mathematical description of this process can be given in terms of a Wick rotation. Time is expressed using imaginary numbers, introduced only for computational convenience. The distinction between the dimensions of time and space disappears. This description is very useful, since it provides a convenient way to determine tunneling probabilities. The most probable universes are the ones with

the smallest initial size and the highest vacuum energy. Once a universe is formed, it immediately starts expanding due to the high energy of the vacuum.

This provides a beginning for the story of eternal inflation.

One might imagine that closed universes are popping out of nothing like bubbles in a glass of champagne, but this analogy would not be quite accurate. Bubbles pop out in liquid, but in the case of universes, there is no space out of which they might pop. A nucleated closed universe is all the space there is, aside from the disconnected spaces of other closed universes. Beyond it, there is no space, and no time.

What causes the universe to pop out of nothing? No cause is needed. If you have a radioactive atom, it will decay, and quantum mechanics gives the decay probability in a given interval of time, say, a minute. There is no reason why the atom decayed at this particular moment and not another. The process is completely random. No cause is needed for the quantum creation of the universe.

The theory of quantum creation is no more than a speculative hypothesis. It is unclear how, or whether, it can be tested observationally. It is nonetheless the first attempt to formulate the problem of cosmic origin and to address it in a quantitative way.[19]

An unaddressable mystery

The answer to the question, "Did the universe have a beginning?" is, "It probably did." We have no viable models of an eternal universe. The BGV theorem gives us reason to believe that such models simply cannot be constructed.

When physicists or theologians ask me about the BGV theorem, I am happy to oblige. But my own view is that the theorem does not tell us anything about the existence of God. A deep mystery remains. The laws of physics that describe the quantum creation of the universe also describe its evolution. This seems to suggest that they have some independent existence.

What exactly this means, we don't know.

And why are these laws the ones we have? Why not other laws?

We have no way to begin to address this mystery.

Appendix: Mathematical details

In this Appendix, I outline a proof of the BGV theorem.

Start with a homogeneous, isotropic, and spatially flat universe with the metric:

$$ds^2 = dt^2 - a^2(t)d\vec{x}^2.$$

The Hubble expansion rate is $H = \dot{a}/a$, where the dot denotes a derivative with respect to time *t*. We can imagine that the universe is filled with comoving particles, moving along the timelike geodesics $\vec{x} = const$ Consider an inertial observer, whose

world line is $x^{\mu}(\tau)$ parametrized by the proper time τ. For an observer of mass m, the 4-momentum is $P^{\mu} = m\, dx^{\mu} / d\tau$ so that $d\tau = (m / E)dt$ where $E = P^0 = \sqrt{p^2 + m^2}$ denotes the energy, and p, the magnitude of the 3-momentum. It follows from the geodesic equation of motion that $p \propto 1/a(t)$ so that $p(t) = [a(t_f) / a(t)]p_f$ where p_f designates the momentum at some reference time t_f

Thus:

$$\int_{t_i}^{t_f} H(\tau)d\tau = \int_{a(t_i)}^{a(t_f)} \frac{m\, da}{\sqrt{m^2a^2 + p_f^2 a^2(t_f)}} = F(\gamma_f) - F(\gamma_i) \leq F(\gamma_f),$$

where $t_i < t_f$ is some initial moment.

Note that:

$$F(\gamma) = \frac{1}{2}\ln\left(\frac{\gamma + 1}{\gamma - 1}\right),$$

where $\gamma = 1/\sqrt{1 - v_{rel}^2}$ is the Lorentz factor, and $v_{rel} = p/E$ is the observer's speed relative to the comoving particles.

For any non-comoving observer, $\gamma > 1$ and $F(\gamma) > 0$.

The expansion rate averaged over the observer's world line is:

$$H_{av} = \frac{1}{\tau_f - \tau_i}\int_{t_i}^{t_f} H(\tau)d\tau.$$

Assuming that

$$H_{av} > 0,$$

and using the first equation, it follows that

$$\tau_f - \tau_i \leq \frac{F(\gamma_f)}{H_{av}}.$$

This implies that any non-comoving past-directed timelike geodesic satisfying the condition $H_{av} > 0$ must have a finite proper length, and so must be past-incomplete.

There is no appealing to homogeneity and isotropy in an arbitrary spacetime. Imagine that the universe is filled with a congruence of comoving geodesics, representing test particles and consider a non-comoving geodesic observer described by a world line $x^{\mu}(\tau)$[20] Let u^{μ} and v^{μ} designate the 4-velocities of test particles and the observer.

Then the Lorentz factor of the observer relative to the particles is

$$\gamma = u_{\mu}v^{\mu}.$$

To characterize the expansion rate in general space-time, it suffices to focus on test particle geodesics that cross the observer's world line.

Consider two such geodesics encountering the observer at times τ and $\tau + \Delta\tau$.

Define the parameter

$$H = \lim_{\Delta\tau \to 0} \frac{\Delta u_r}{\Delta r},$$

where Δu_r is the relative velocity of the particles in the observer's direction of motion, and Δr is the particle's separation. Both quantities are computed in the rest frame of one of the particles.

For a homogeneous and isotropic universe, this definition reduces to the Hubble parameter.

The expansion parameter can be expressed as a total derivative,

$$H = \frac{d}{d\tau} F\big(\gamma(\tau)\big).$$

The integral of H along the observer's world line is still given by the difference of $F(\gamma)$ at its endpoints. Conclusions about homogeneous and isotropic universes carry over immediately to generic universes.

There remains the null observer described by a null geodesic. The role of the proper time τ is then played by an affine parameter. BGV showed that, with a suitable normalization of τ, the expansion rate is given by

$$H = \frac{d}{d\tau} F\big(\gamma(\tau)\big),$$

with $F(\gamma) = 1/\gamma$, and γ defined by

$$H = \lim_{\Delta\tau \to 0} \frac{\Delta u_r}{\Delta r}.$$

Clearly, $F(\gamma) > 0$, and the argument goes through as before.

A rigorous formulation of the BGV theorem is now possible. Let λ be a timelike or null geodesic maximally extended to the past, and let C be a timelike geodesic congruence defined along λ.

If the expansion rate of C averaged along λ is positive, then λ. must be past-incomplete.

Notes

1 See, for example, Viatcheslav Mukhanov, *Physical Foundations of Cosmology* (Cambridge: University Press Cambridge, 2005).

2 In general relativity, space and time are unified in a single four-dimensional entity, called space-time.

3 Roger Penrose, "Gravitational Collapse and Space-time Singularities," *Physical Review Letters* 14, no. 57 (1965): 57–59.
4 Mathematically, the null convergence condition requires that the Ricci curvature tensor $R_{\mu\nu}$ must satisfy $R_{\mu\nu}N^{\mu}N^{\nu} \geq 0$ for all null vectors N^{μ}. (A null vector is a vector of zero norm, $N^{\mu}N_{\mu} = 0$.)
5 Lawrence Ford, "Negative Energy Densities in Quantum Field Theory," *International Journal of Modern Physics A* A25, no. 2355 (2010).
6 For a review of cosmic inflation, see Alan Guth, *The Inflationary Universe* (New York: Basic Books, 1998); Alexander Vilenkin, *Many Worlds in One* (New York: Hill and Wang, 2006).
7 This is not excluded by Penrose's singularity theorem, since the null convergence condition that Penrose used in the proof is generally violated by quantum fluctuations in a false vacuum.
8 Arvind Borde, Alan Guth and Alexander Vilenkin, "Inflationary Spacetimes are Incomplete in Past Directions," *Physical Review Letters* 90 (2003): 1–4.
9 George Ellis, Jeff Murugan and Christos Tsagas, "The Emergent Universe: An Explicit Construction," *Classical and Quantum Gravity* 21 (2004): 233–50.
10 Audrey Mithani and Alexander Vilenkin, "Instability of an Emergent Universe," *Journal of Cosmology and Astroparticle Physics* 1405 (2014).
11 Paul Steinhardt and Neil Turok, "A Cyclic Model of the Universe," *Science* 296 (2002): 1,436–39.
12 For a uniform (on average) distribution of matter, entropy is proportional to volume; if the volume grows without bounds, entropy is also unbounded.
13 Edward Milne, *Modern Cosmology and the Christian Idea of God* (Oxford: Clarendon Press, 1952).
14 As quoted in Carl Friedrich von Weizsäcker, *The Relevance of Science: Creation and Cosmology* (New York: Harper, 1964).
15 "Daniel Dennett on William Lane Craig," YouTube (August 12, 2009); "William Lane Craig on Daniel Dennett," YouTube (December 14, 2012); "Dawkins Vs Creationist," YouTube (January 15, 2014).
16 William Craig, "Theistic Critiques of Atheism," in *The Cambridge Companion to Atheism*, ed. Michael Martin (Cambridge: Cambridge University Press, 2007), 69–85.
17 This can be illustrated by the following Newtonian example. Consider a system of two masses interacting gravitationally. The total energy of the system is the sum of the kinetic energies of their motion and the gravitational energy. (I disregard the rest energy of the masses, which remains constant for motion much slower than the speed of light.) Suppose the masses start out at rest at a large distance from one another. Both kinetic and gravitational energies are then equal to zero, so the total energy is zero. As the masses accelerate towards one another under the action of gravity, their kinetic energy grows, and by the law of energy conservation, gravitational energy also grows, but with an opposite sign.
18 The idea of the quantum creation of universes from nothing was suggested in my 1982 paper, and later, in a somewhat different form, by James Hartle and Stephen Hawking, and also by Andrei Linde. See Alexander Vilenkin, "Creation of Universes from Nothing," *Physics Letters B* 117 (1982): 25–28; James Hartle and Stephen Hawking, "Wave Function of the Universe," *Physical Review D* 28, no. 12 (1983): 2,960–75; Andrei Linde, "Quantum Creation of the Inflationary Universe," *Lettere al Nuovo Cimento* 39, no. 17 (1984): 401–5.
19 David Albert, "On the Origin of Everything," *New York Times*, March 23, 2012.
20 In fact, the geodesic congruence needs to be defined only along the observer's world line.

5

A Dying Universe: The Long-Term Fate and Evolution of Astrophysical Objects

Fred C. Adams and Gregory Laughlin

I Introduction

The long-term future of the universe and its contents is a topic of profound scientific and philosophical importance. With our current understanding of physics and astrophysics, many of the questions regarding the ultimate fate of the universe can now be quantitatively addressed. Our goal is to summarize and continue the development of a quantitative theory of the future.

Investigations of the early universe at both accessible and inaccessible energies have become commonplace, and a great deal of progress within this discipline has been made (see, e.g., Weinberg, 1972, 1977; Kolb & Turner, 1990; Linde, 1990; Peebles, 1993; Zuckerman & Malkan, 1996). On the other hand, relatively little work has focused on the future of the universe. The details of the fiery denouement in store for a closed universe have been outlined by Rees (1969), whereas an overview of the seemingly more likely scenario in which the universe is either open or flat, and hence expands forever, was set forth in the seminal paper *Time Without End* (Dyson, 1979). The development of an open universe was also considered in detail by Islam (1977, 1979). The spirit of Rees, Islam, and Dyson's work inspired several follow-up studies (see also Rees, 1981). The forthcoming evolution of very low-mass stars has been discussed in general terms by Salpeter (1982). The effects of matter annihilation in the late universe were studied (Page & McKee, 1981ab), and some aspects of proton decay have been explored (Dicus et al., 1982; Turner, 1983). Finally, the possibility of self-reproducing inflationary domains has been proposed (Linde, 1988). In general, however, the future of the universe has not been extensively probed with rigorous calculations.

Because the future of the universe holds a great deal of intrinsic interest, a number of recent popular books have addressed the subject (e.g., Davies, 1994; Dyson, 1988; Barrow & Tipler, 1986; Poundstone, 1985). Authors have also grappled with the intriguing prospects for continued life, both human and otherwise, in far future (e.g., Dyson, 1979; Frautschi, 1982; Barrow & Tipler, 1986; Linde, 1988, 1989; Tipler, 1992; Gott, 1993; Ellis & Coule, 1994). Our aim, however, is to proceed in as quantitative a manner as possible. We apply known physical principles to investigate

the future of the universe on planetary, stellar, galactic, and cosmic scales. The issue of life, however alluring, is not considered here.

In standard Big Bang Cosmology, evolutionary epochs are usually expressed in terms of the redshift. When considering the far future, however, time itself is often the more relevant evolutionary measure. The immense dynamic range of time scales τ involved in the subject suggests a convenient logarithmic unit of time η, defined by

$$\eta \equiv \log_{10}\left[\frac{\tau}{1\text{yr}}\right]. \tag{1.1}$$

We refer to a particular integer value of η as a "cosmological decade". For example, the current age of the universe corresponds to $\eta \approx 10$.

The article of faith inherent in our discussion is that the laws of physics are constant in time, at least over the range of time scales $10 < \eta < 100$ under consideration. There is no general guarantee that this assumption holds. Nevertheless, modern cosmology suggests that physical laws have held constant from the Planck time to the present, i.e., over cosmological decades spanning the range $-50 \leq \eta \leq 10$, and there is little reason to expect that they will not continue to do so. We also implicitly assume that all of the relevant physics is known (with full awareness of the fact that our version of the future will be subject to revision as physical understanding improves).

This paper is organized in roughly chronological order, moving from events in the relatively near future to events in the far future. In section §II, we discuss physical processes that affect conventional stellar evolution; these processes will take place in the time range $10 < \eta < 15$. In §III, we discuss events which lead to the disruption and death of the galaxy; these processes unfold over a time range $15 < \eta < 25$. Marching further into time, in §IV, we discuss the fate of stellar objects in the face of very long-term processes, including proton decay ($30 < \eta < 40$), and Hawking radiation ($60 < \eta < 100$). In §V, we broaden our scope and focus on the long-term evolution of the universe as a whole. We conclude, in §VI, with a general overview of our results. Since physical eschatology remains embryonic, we emphasize the major unresolved issues and point out possible avenues for further research.

II The end of conventional stellar evolution

At the present epoch, stars are the cornerstone of astrophysics. Stars mediate the appearance and evolution of galaxies, stars are responsible for evolving the chemical composition of matter, and stars provide us with much of the information we have regarding the current state of the universe.

For the next several thousand Hubble times, conventionally evolving stars will continue to play the central role. We thus consider the forthcoming aspects of our current epoch, which we term the *Stelliferous Era*. In particular, the fact that the majority of stars have barely begun to evolve motivates an extension of standard stellar evolution calculations of very low-mass stars to time scales much longer than

the current age of the universe. We also discuss continued star formation within the galaxy, and the final mass distribution of stellar remnants.

A Lifetimes of main-sequence stars

Low-mass stars are by far the most commonplace (e.g., Henry, Kirkpatrick, & Simons, 1994), and they live for a long time. To a working approximation, the main-sequence (core-hydrogen-burning) lifetime of a star depends on its mass through the relation

$$\tau_* = 10^{10}\,\mathrm{yr}\left[\frac{M_*}{1M_\odot}\right]^{-\alpha}, \tag{2.1a}$$

where the index $\alpha \approx 3 - 4$ for stars of low mass. In terms of cosmological decades η, we obtain

$$\eta_* = 10 - \alpha \log_{10}\left[M_* / 1M_\odot\right]. \tag{2.1b}$$

Thus, for example, $\eta_* \approx 13$ for a small star with $M_* = 0.1\ \mathrm{M}_e$. Indeed at the present time, only stars with masses $M_* > 0.8\mathrm{M}_e$ have had time to experience significant post-main-sequence evolution. Hence, a large fraction,

$$f \equiv \frac{\int_{M_{min}}^{0.8} \left(dN/dm\right) dm}{\int_{M_{min}}^{M_{max}} \left(dN/dm\right) dm} \sim 80\% \tag{2.2}$$

of all stars ever formed have yet to experience any significant evolution (here, *dN/dm* is the mass distribution – see §II.E). We are effectively still in the midst of the transient initial phases of the stelliferous epoch.

Very little consideration has been given to the post-main-sequence development of stars which are small enough to outlive the current age of the universe. An essay by Salpeter (1982) contains a qualitative discussion regarding the evolution of M stars (especially with respect to ^{3}He production) but detailed stellar evolutionary sequences have not been presented in the literature. Nevertheless, there is a sizable collection of papers which discuss the pre-main-sequence and main-sequence properties of very low-mass stars (e.g., Kumar, 1963; Copeland, Jensen & Jorgensen, 1970; Grossman & Graboske, 1971; D'Antona & Mazzitelli, 1985; Dorman, Nelson & Chau, 1989). The best comprehensive family of models spanning the M dwarfs and brown dwarfs is probably that of Burrows et al. (1993). Those authors devote attention to the formative cooling phases, as well as the exact mass of the minimum mass star (which for their input physics occurs at $M_* = 0.0767\ M_e$). Evolution beyond 20 billion years was not considered (see also Burrows & Liebert, 1993).

The dearth of information regarding the fate of the M dwarfs has recently been addressed (Laughlin, Bodenheimer, & Adams, 1996). We have performed a detailed series of stellar evolution calculations which follow the pre-main-sequence through

post-main-sequence evolution of late M-dwarfs, yielding the following picture of what lies in store for the low-mass stars.

Newly formed stars containing less mass than $M_* \sim 0.25M_e$ are fully convective throughout the bulk of their structure. The capacity of these stars to entirely mix their contents has several important consequences. First, these late M stars maintain access to their entire initial reserve of hydrogen, greatly extending their lifetimes in comparison to heavier stars like the sun which see their fuel supply constricted by stratified radiative cores. Second, as recognized by Salpeter (1982), full convection precludes the buildup of composition gradients which are ultimately responsible (in part) for a star's ascent up the red giant branch. The lowest-mass stars burn all their hydrogen into helium over an $\eta = 13$ time scale, and then quietly fade from prominence as helium white dwarfs. This general evolutionary scenario is detailed in Figure 5.1 (adapted from Laughlin et al., 1996), which charts the path in the Hertzsprung-Russell diagram followed by low-mass stars of several different masses in the range $0.08M_e \leq M_* \leq 0.25M_e$.

Upon emerging from its parent cloud core, the lowest-mass star capable of burning hydrogen ($M_* \approx 0.08M_e$) descends the convective Hayashi track and arrives on the main-sequence with a luminosity $L_* \sim 10^{-4}L_e$. The main-sequence phase is characterized by gradual prolonged increase in both luminosity and effective surface temperature as hydrogen is consumed. Due to the relatively low prevailing temperature in the stellar core ($T_c \approx 4 \times 10^6$ K), the proton-proton nuclear reaction chain is decoupled from statistical equilibrium, and the concentration of ^{3}He increases steadily until $\eta = 12.6$, at which time a maximum-mass fraction of 16% ^{3}He has been attained. As the initial supply of hydrogen is depleted, the star heats up and contracts, burns the ^{3}He, increases

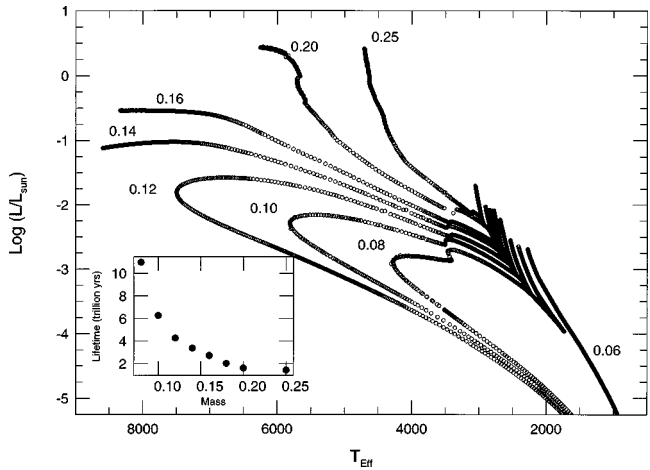

Figure 5.1. The Hertzsprung-Russell diagram for low-mass stars for time scales much longer than the current age of the universe. The labeled curves show the evolutionary tracks for stars of varying masses, from 0.08 $M_\odot$ to 0.25$M_\odot$, as well as the brown dwarf track for a substellar object with mass $M_* = 0.06M_\odot$. The inset figure shows the main-sequence lifetimes as a function of stellar mass.

in luminosity by a factor of 10, and more than doubles its effective temperature. After ~11 trillion years, when the star has become 90% ^{4}He by mass, a radiative core finally develops. The evolutionary time scale begins to accelerate, and hydrogen is exhausted relatively quickly in the center of the star. When nuclear burning within the modest resulting shell source can no longer provide the star's mounting energy requirements, the star begins to contract and cool and eventually becomes a helium white dwarf. Stars with masses up to ~0.20 M_e follow essentially this same evolutionary scenario. As stellar mass increases, radiative cores develop sooner, and the stars perform increasingly dramatic blueward excursions in the H-R diagram.

A star with a slightly larger mass, $M_* = 0.23\ M_e$, experiences the onset of a radiative core when the hydrogen mass fraction dips below 50%. The composition gradients which ensue are sufficient to briefly drive the star to lower effective temperature as the luminosity increases. In this sense, stars with mass $M_* = 0.23 M_e$ represent the lowest-mass objects that can become conventional "Red Giants". At these low masses, however, the full giant phase is not completed. Stars with initial mass $M_* < 0.5\ M_e$ will be unable to generate the high central temperatures ($T_c \sim 10^8$ K) required for the helium flash; these stars abort their ascent up the giant branch by veering to the left in the H-R diagram in the manner suggested by Figure 5.1.

The steady luminosity increases experienced by aging M dwarfs will have a considerable effect on the mass-to-light ratio of the galaxy. For example, as a 0.2 M_e star evolves, there is a relatively fleeting epoch (at $\eta \approx 12$) during which the star has approximately the same radius and luminosity as the present-day sun. Given that M dwarfs constitute the major fraction of all stars, the total luminosity of the galaxy will remain respectably large, $L_{gal} \sim 10^{10}\ L_e$, at this future date. This luminosity is roughly comparable to the characteristic luminosity $L^* = 3.4 \times 10^{10}\ L_e$ displayed by present-day galaxies (Mihalas & Binney, 1981).

B Forthcoming metallicity effects

The foregoing evolutionary calculations assumed a solar abundance set. In the future, the metallicity of the galaxy will steadily increase as stars continue to process hydrogen and helium into heavy elements. It is thus useful to determine the effects of these metallicity increases.

1 Stellar lifetimes vs metallicity

First, it is possible to construct a simple scaling relation that clarifies how stellar lifetimes τ_* depend on the metallicity Z. The stellar lifetime is roughly given by amount of fuel available divided by the rate of fuel consumption, i.e.,

$$\tau_* \sim M_* X / L, \tag{2.3}$$

where M_* is the stellar mass and X is the hydrogen mass fraction. For relatively low-mass stars, the luminosity L obeys the scaling relation

$$L \sim k_0^{-} \mu^{7.5} M_*^{5.5}, \tag{2.4}$$

where μ is the mean molecular weight of the star and where κ_0 is the constant of proportionality appearing in the usual opacity relation for stars (Clayton, 1983). Thus, for a given stellar mass M_*, the lifetime scales according to

$$\tau_* \sim k_0 X \mu^{-7.5}. \tag{2.5}$$

To evaluate the stellar lifetime scaling relation, one needs to know how the parameters κ_0, X, and μ vary with metallicity. The opacity constant κ_0 is roughly linearly dependent on the metallicity, i.e.,

$$k_0 \sim Z. \tag{2.6}$$

The mean molecular weight μ can be approximately written in the form

$$\mu \approx \frac{2}{\left(1+3X+Y/2\right)}, \tag{2.7}$$

where Y is the helium mass fraction (e.g., see Clayton, 1983). By definition, the mass fractions obey the relation

$$X+Y+Z=1. \tag{2.8}$$

Finally, for this simple model, we write the helium abundance Y in the form

$$Y = Y_P + fZ. \tag{2.9}$$

where Y_P is the primordial abundance and the factor f accounts for the increase in helium abundance as the metallicity increases. Big Bang nucleosynthesis considerations indicate that $Y_P \approx 1/4$ (Kolb & Turner, 1990), whereas $f \approx 2$ based on the solar enrichment in Y and Z relative to the primordial values. Combining the above results, we obtain a scaling relation for the dependence of stellar lifetimes on metallicity,

$$\tau_* \sim Z\left(1-aZ\right)\left(1-bZ\right)^{7.5}, \tag{2.10}$$

where we have defined constants $a \equiv 4(1 + f)/3 \approx 4$ and $b \equiv 8/9 + 20f/27 \approx 64/27$. This result implies that stellar lifetimes have a *maximum value*. In particular, we find that stars born with metallicity $Z \approx 0.04$ live the longest. For larger values of Z, the reduction in nuclear fuel and the change in composition outweigh the lifetime extending decrease in luminosity arising from the increased opacity.

A recent set of galactic chemical evolution calculations (Timmes, 1996) have probed far into the stelliferous epoch. The best indications suggest that the galactic abundance set will approach an asymptotically constant composition ($X \sim 0.2$, $Y \sim 0.6$, and $Z \sim 0.2$)

over a time scale $\eta \sim 12$. As a consequence, any generations of stars formed after $\eta \sim 12$ will suffer significantly shorter lifetimes than the theoretical maximum implied by equation [2.10].

2 Stellar masses vs metallicity

The maximum stable stellar mass decreases as metallicity increases. On the main sequence, the maximum possible mass is reached when the star's radiation pressure comes to dominate the thermal (gas) pressure within the star. Here, we introduce the usual ansatz that the total pressure at the center of the star can be written in the form $P_C = P_R + P_g$, where the thermal gas pressure is given by the fraction $P_g = \beta P_C$ and, similarly, $P_R = (1 - \beta)P_C$. Using the ideal gas law for the thermal pressure and the equation of state for a gas of photons, we can write the central pressure in the form

$$P_C = \left[\frac{3}{a}\frac{(1-\beta)}{\beta^4}\right]^{1/3}\left[\frac{k\rho_C}{\mu m_P}\right]^{4/3}, \tag{2.11}$$

where k is Boltzmann constant and a is the radiation constant. The quantity μ is again the mean molecular weight and can be written in the form of equation [2.7]. In hydrostatic equilibrium, the central pressure required to support a star of mass $M_\star$ can be expressed as

$$P_C \approx \left[\frac{\pi}{36}\right]^{1/3} GM_\star^{2/3}\rho_C^{4/3}, \tag{2.12}$$

where ρ_C is the central density (see Phillips, 1994).

Equating the above two expressions [2.11] and [2.12], we can solve for the mass to find

$$M_\star = \left[\frac{108}{\pi a}\frac{1-\beta}{\beta^4}\right]^{1/2}\left[\frac{k}{\mu m_P}\right]^2 G^{-3/2} \approx 40M_\odot\mu^{-2}, \tag{2.13}$$

where we have set $\beta = 1/2$ to obtain the numerical value. The maximum stellar mass thus depends somewhat sensitively on the mean molecular weight μ, which in turn is a function of the metallicity. By applying the approximations [2.7], [2.8], and [2.9], one can write the maximum mass in the form

$$M_\star = 40M_\odot\left\{\left(2-5Y_P/4\right)-\left(3+5f/2\right)Z/2\right\}^2 \approx 144M_\odot\left(1-2.4Z\right)^2. \tag{2.14}$$

Thus, for the expected asymptotic value of the metallicity, $Z = 0.2$, the maximum-mass star is only $M_\star \approx 30M_\odot$.

The continuously increasing metallicity of the interstellar medium will also have implications for low-mass stars. Higher metallicity leads to more effective cooling, which leads to lower temperatures, which in turn favors the formation of less massive

stars (e.g., see the recent theory of the initial mass function by Adams & Fatuzzo, 1996). The IMF of the future should be skewed even more dramatically in favor of the faintest stars.

The forthcoming metallicity increases may also decrease the mass of the minimum-mass main-sequence star as a result of opacity effects (cf. the reviews of Stevenson, 1991; Burrows & Liebert, 1993). Other unexpected effects may also occur. For example, when the metallicity reaches several times the solar value, objects with mass $M_\star = 0.04\ M_e$ may quite possibly halt their cooling and contraction and land on the main sequence when thick ice clouds form in their atmospheres. Such "frozen stars" would have an effective temperature of $T_\star \approx 273$ K, far cooler than the current minimum-mass main-sequence stars. The luminosity of these frugal objects would be more than a thousand times smaller than the dimmest stars of today, with commensurate increases in longevity.

C The fate of the Earth and the Sun

A popular and frequently quoted scenario for the demise of the Earth involves destruction through evaporation during the Sun's asymptotic giant branch (AGB) phase. As the Sun leaves the horizontal branch and expands to become an AGB star, its outer radius may swell to such an extent that the photospheric radius overtakes the current orbital radius of the Earth. If this state of affairs comes to pass, then two important processes will affect the Earth: [1] Evaporation of material due to the extreme heat, and [2] Orbital decay through frictional drag. This second process drives the Earth inexorably into the giant sun, thereby increasing the efficacy of the evaporation process. Once the earth finds itself *inside* the sun, the time scale for orbital decay is roughly given by the time required for the expiring Earth to sweep through its mass, M_E, in solar material. This short time interval is given by

$$\tau = \frac{M_E}{\rho_\odot \left(\pi R_E^2\right) \upsilon_{\text{orbit}}} \approx 50\text{yr}, \tag{2.15}$$

where $\rho_e \sim 10^{-6}$ g/cm^3 is the mass density of solar material at the photosphere, $R_E \approx 6370$ km is the radius of the Earth, and $v_{orbit} \approx 30$ km/s is the orbital speed. Hence, the demise of the Earth will befall it swiftly, even in comparison to the accelerated stellar evolution time scale inherent to the asymptotic giant branch. The Earth will be efficiently dragged far inside the sun and vaporized in the fierce heat of the stellar plasma, its sole legacy being a small (0.01%) increase in the metallicity of the Sun's distended outer envelope.

Recent work suggests, however, that this dramatic scene can be avoided. When the sun reaches a luminosity of $\sim 100 L_e$ on its ascent of the red giant branch, it will experience heavy mass loss through the action of strong stellar winds. Mass loss results in an increase in the orbital radii of the planets and can help the Earth avoid destruction. However, the actual amount of mass loss remains uncertain; estimates are based largely on empirical measurements (see Reimers, 1975), but it seems reasonable

that the sun will diminish to ~0.70M_e when it reaches the tip of the red giant branch, and will end its AGB phase as a carbon white dwarf with mass ~0.5M_e. Detailed stellar evolution calculations for the sun have been made by Sackmann, Boothryod, & Kraemer (1993). In their best-guess mass-loss scenario, they find that the orbital radii for both the Earth and Venus increase sufficiently to avoid being engulfed during the AGB phase. Only with a more conservative mass-loss assumption, in which the Sun retains 0.83M_e upon arrival on the horizontal branch, does the solar radius eventually overtake the Earth's orbit.

D Continued star formation in the galaxy

Galaxies can only live as long as their stars. Hence it is useful to estimate how long a galaxy can sustain normal star formation (see, e.g., Shu, Adams, & Lizano, 1987) before it runs out of raw material. One would particularly like to know when the *last* star forms.

There have been many studies of the star formation history in both our galaxy as well as other disk galaxies (e.g., Roberts, 1963; Larson & Tinsley, 1978; Rana, 1991; Kennicutt, Tamblyn, & Congdon, 1994; hereafter KTC). Although many uncertainties arise in these investigations, the results can be roughly summarized as follows. The gas depletion time τ_R for a disk galaxy is defined to be the current mass in gas, M_{gas}, divided by the star formation rate *SFR,* i.e.,

$$\tau_R \equiv \frac{M_{\rm gas}}{SFR} \,. \tag{2.16}$$

For typical disk galaxies, this time scale is comparable to the current age of the universe; KTC cite a range $\tau_R \approx 5 - 15$ Gyr. The actual time scale for (total) gas depletion will be longer because the star formation rate is expected to decrease as the mass in gas decreases. For example, if we assume that the star formation rate is proportional to the current mass in gas, we derive a total depletion time of the form

$$\tau = \tau_R \ln\left[M_0/M_F \right] \tag{2.17}$$

where M_0 is the initial mass in gas and M_F is the final mass. For typical disk galaxies, the initial gas mass is $M_0 \sim 10^{10} M_e$ (see Table 5 of KTC). Thus, if we take the extreme case of $M_F = 1\ M_e$, the total gas depletion time is only $\tau \approx 23\tau_R \approx 120 - 350$ Gyr. In terms of cosmological decades, the gas depletion time becomes $\eta_D = 11.1 - 11.5$.

Several effects tend to extend the gas depletion time scale beyond this simple estimate. When stars die, they return a fraction of their mass back to the interstellar medium. This gas recycling effect can prolong the gas depletion time scale by a factor of 3 or 4 (KTC). Additional gas can be added to the galaxy through infall onto the galactic disk, but this effect should be relatively small (cf. the review of Rana, 1991); the total mass added to the disk should not increase the time scale by more than a factor of 2. Finally, if the star formation rate decreases more quickly with decreasing

gas mass than the simple linear law used above, then the depletion time scale becomes correspondingly larger. Given these complications, we expect the actual gas depletion time will fall in the range

$$\eta_D = 12 - 14 \tag{2.18}$$

Thus, by the cosmological decade $\eta \approx 14$, essentially all normal star formation in galaxies will have ceased. Coincidentally, low-mass M dwarfs have life expectancies that are comparable to this time scale. In other words, both star formation and stellar evolution come to an end at approximately the same cosmological decade.

There are some indications that star formation may turn off even more dramatically than outlined above. Once the gas density drops below a critical surface density, star formation may turn off completely (as in elliptical and S0 galaxies). The gas may be heated entirely by its slow accretion onto a central black hole.

These results indicate that stellar evolution is confined to a reasonably narrow range of cosmological decades. It is presumably impossible for stars to form and burn hydrogen before the epoch of recombination in the universe (at a redshift $z \sim 1000$ and hence $\eta \sim 5.5$). Thus, significant numbers of stars will exist only within the range

$$5.5 < \eta < 14. \tag{2.19}$$

The current epoch ($\eta \sim 10$) lies near the center of this range of (logarithmic) time scales. On the other hand, if we use a linear time scale, the current epoch lies very near the beginning of the stelliferous era.

E The final mass function

When ordinary star formation and conventional stellar evolution have ceased, all of the remaining stellar objects will be in the form of brown dwarfs, white dwarfs, neutron stars, and black holes. One way to characterize the stellar content of the universe at this epoch is by the mass distribution of these objects; we refer to this distribution as the "Final Mass Function" or FMF. Technically, the Final Mass Function is not final in the sense that degenerate objects can also evolve and thereby change their masses, albeit on vastly longer time scales. The subsequent evolution of degenerate objects is discussed in detail in §IV.

Two factors act to determine the FMF: [1] The initial distribution of stellar masses (the initial mass function [IMF] for the progenitor stars), and [2] The transformation between initial stellar mass and the mass of the final degenerate object. Both of these components can depend on cosmological time. In particular, one expects that metallicity effects will tend to shift the IMF toward lower masses as time progresses.

The initial mass function can be specified in terms of a general log-normal form for the mass distribution $\psi = dN/d \ln m$,

$$\ln\psi\left(\ln m\right) = A - \frac{1}{2\langle\alpha\rangle^2}\left\{\ln\left[m / m_C\right]\right\}^2, \tag{2.20}$$

where A, m_C, and $\langle\sigma\rangle$ are constants. Throughout this discussion, stellar masses are written in solar units, i.e., $m \ldots M_*/(1M_e)$. This general form for the IMF is motivated by the both current theory of star formation and by general statistical considerations (Adams & Fatuzzo, 1996; Zinnecker, 1984; Larson, 1973; Elmegreen & Mathieu, 1983). In addition, this form is (roughly) consistent with observations (Miller & Scalo, 1979), which suggest that the shape parameters have the values $\langle\sigma\rangle \approx 1.57$ and $m_C \approx 0.1$ for the present-day IMF (see also Salpeter, 1955; Scalo, 1986; Rana, 1991). The constant A sets the overall normalization of the distribution and is not of interest here.

For a given initial mass function, we must find the final masses m_F of the degenerate objects resulting from the progenitor stars with a given mass m. For the brown dwarf range of progenitor masses, $m < m_H$, stellar objects do not evolve through nuclear processes and hence $m_F = m$. Here, the scale $m_H \approx 0.08$ is the minimum stellar mass required for hydrogen burning to take place.

Progenitor stars in the mass range $m_H \leq m \leq m_{SN}$ eventually become white dwarfs, where the mass scale $m_{SN} \approx 8$ is the minimum stellar mass required for the star to explode in a supernova (note that the mass scale m_{SN} can depend on the metallicity – see Jura, 1986). Thus, for the white dwarf portion of the population, we must specify the transformation between progenitor mass m and white dwarf mass m_{WD}. The results of Laughlin et al. (1996) indicate that stars with main sequence masses $m < 0.4$ will undergo negligible mass-loss in becoming helium white dwarfs. Unfortunately, this relationship remains somewhat ill-defined at higher masses, mostly due to uncertainties in red giant mass-loss rates (e.g., see Wood, 1992). For the sake of definiteness, we adopt the following transformation between progenitor mass and white dwarf mass,

$$m_{WD} = \frac{m}{1+\alpha m}\exp\left[\beta m\right], \tag{2.21}$$

with $\alpha = 1.4$ and $\beta = 1/15$. This formula is consistent with the models of Wood (1992) over the appropriate mass range and approaches the expected form $m_{WD} = m$ in the low-mass limit.

Stars with large initial masses, $m > m_{SN}$, end their lives in supernova explosions and leave behind a neutron star (although black holes can also, in principle, be produced). The mass of the remnant neutron star is expected to be near the Chandrasekhar limit $m_{Ch} \approx 1.4$, as confirmed in the case of the binary pulsar (Manchester & Taylor, 1977).

To compute the FMF, one convolves the initial mass function with the transformations from progenitor stars to white dwarfs and neutron stars. The Final Mass Function that results is shown in Figure 5.2. For comparison, the initial mass function is also shown (as the dashed curve). Notice that the two distributions are similar for masses less than the Chandrasekhar mass (~$1.4M_e$) and completely different for larger masses.

Once the FMF has been determined, one can estimate the number and mass fractions of the various FMF constituents. We define N_{BD} to be the fraction of brown

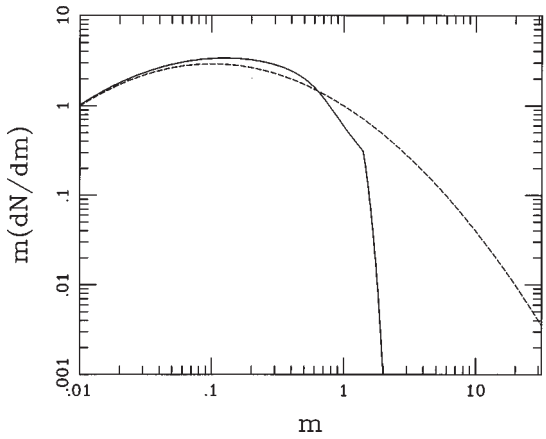

Figure 5.2 The Final Mass Function (FMF) for stars. Solid curve shows the predicted distribution m(dN/dm) for the masses of the degenerate stellar objects (brown dwarfs, white dwarfs, and neutron stars) remaining at the cosmological epoch when conventional star formation has ceased. The dashed curve shows the mass distribution of the initial progenitor population (the initial mass function).

dwarfs by number and M_{BD} to be the fraction of brown dwarfs by mass, with analogous fractions for white dwarfs (N_{WD} and M_{WD}) and neutron stars (N_{NS} and M_{NS}). For an IMF of the form [2.20] with present-day values for the shape parameters, we obtain the following number fractions:

$$N_{BD} = 0.45, \quad N_{WD} = 0.55, \quad N_{NS} = 0.0026. \tag{2.22}$$

Similarly, for the mass fractions one finds

$$M_{BD} = 0.097, \quad M_{WD} = 0.88, \quad M_{NS} = 0.024. \tag{2.23}$$

Thus, brown dwarfs are expected to be present in substantial numbers, but most of the mass will reside in the form of white dwarfs. Neutron stars will make a relatively small contribution to the total stellar population. The above values for N_{NS}, and M_{NS} were obtained under the assumption that all stars $m > m_{SN} \sim 8$ produce neutron stars. In reality, a portion of these high mass stars may collapse to form black holes instead, but this complication does not materially affect the basic picture described above.

III Death of the galaxy

We have argued that over the long term, the galaxy will incorporate a large fraction of the available baryonic matter into stars. By the cosmological decade $\eta = 14 - 15$,

the stellar component of the galaxy will be in the form of seemingly inert degenerate remnants. Further galactic activity will involve these remnants in phenomena which unfold over time scales ranging from $\eta \sim 15 - 30$. This time period is part of what we term the *Degenerate Era*.

The course of this long-term galactic dynamical evolution is dictated by two generalized competing processes. First, in an isolated physical system containing any type of dissipative mechanism (for example, gravitational radiation, or extremely close inelastic encounters between individual stars), the system must evolve toward a state of lower energy while simultaneously conserving angular momentum. The net outgrowth of this process is a configuration in which most of the mass is concentrated in the center and most of the angular momentum is carried by small parcels at large radii. (The present-day solar system presents a good example of this process at work.) Alternatively, a second competing trend occurs when collisionless relaxation processes are viable. In a galaxy, distant encounters between individual stars are effectively collisionless. Over time, stars tend to be evaporated from the system, the end product of this process is a tightly bound agglomeration (perhaps a massive black hole) in the center, containing only a fairly small fraction of the total mass. Hence, one must estimate the relative efficiencies of both collisionless and dissipative processes in order to predict the final state of the galaxy. This same competition occurs for physical systems on both larger scales (e.g., galaxy clusters) and smaller scales (e.g., globular clusters).

In addition to gravitational radiation and dynamical relaxation, occasional collisions between substellar objects – brown dwarfs – provide a channel for continued star formation at a very slow rate. Collisions and mergers involving two white dwarfs will lead to an occasional type-I supernova, whereas rare impacts involving neutron stars will engender even more exotic bursts of energy. Such events are impressive today. They will be truly spectacular within the cold and impoverished environment of an evolved galaxy.

A Dynamical relaxation of the galaxy

A stellar system such as a galaxy relaxes dynamically because of stellar encounters. The characteristic time scale associated with this process in the case of purely stellar systems is well known and can be written as

$$\tau_{\text{relax}} = \frac{R}{\upsilon} \frac{N}{12 \ln(N/2)}, \tag{3.1}$$

where R is the size of the system, υ is the typical random velocity, and N is the total number of stars (for further discussion, see Lightman & Shapiro, 1978; Shu, 1982; Binney & Tremaine, 1987). The logarithmic factor appearing in the denominator takes into account the effects of many small angle deflections of stars through distant encounters. The time scale for stars to evaporate out of the system is roughly given by

$$\tau_{\text{evap}} = 100 \ \tau_{\text{relax}} \sim 10^{19} \text{yr}, \tag{3.2}$$

where we have used R = 10 kpc, v = 40 km/s, and $N = 10^{11}$ to obtain the numerical result. We thus obtain the corresponding estimate

$$\eta_{\text{evap}} = 19 + \log_{10}\left[R/10\text{kpc}\right] + \log_{10}\left[N/10^{11}\right]. \tag{3.3}$$

Thus, stars escape from the galaxy with a characteristic time scale $\eta \approx 19 - 20$ (see also Islam, 1977; Dyson, 1979).

The stellar dynamical evolution of the Galaxy is more complicated than the simple picture outlined above. First, the galaxy is likely to have an extended halo of dark matter, much of which may be in non-baryonic form. Since this dark halo does not fully participate in the dynamical relaxation process, the halo tends to stabilize the system and makes the stellar evaporation time scale somewhat longer than the simple estimate given above.

Other dynamical issues can also be important. In globular clusters, for example, mass segregation occurs long before stellar evaporation and binary star heating plays an important (actually dominant) role in the long-term evolution. On the other hand, equation [3.1] is formally valid only if the stars are not bound into binary or triple systems. Binary interaction effects can be important for the long-term evolution of the stellar component of the galaxy. In particular, the presence of binaries can increase the effective interaction cross section and can lead to a variety of additional types of interactions. Both three-body interactions and binary-binary interactions are possible. As a general rule, interactions lead to hard binaries becoming harder and wide binaries becoming softer or even disrupted ("ionized"). Binaries that become sufficiently hard (close) can spiral inwards, become mass transfer systems, and eventually explode as supernovae. These effects are just now becoming understood in the context of globular cluster evolution (for further discussion of these dynamical issues, see, e.g., Chernoff & Weinberg, 1990; Hut et al., 1992).

Galaxies in general, and our galaxy in particular, live in groups or clusters. These larger scale systems will also undergo dynamical relaxation processes analogous to those discussed above. However, a more immediate issue that can affect our galaxy in the relatively near future is the possibility of merging with other galaxies in the local group, in particular Andromeda (M31). The orbits of nearby galaxies have been of the subject of much study (e.g., Peebles, 1994), but large uncertainties remain. For the current separation between the Milky Way and M31 (d = 0.75 Mpc) and radial velocity (v_r = 120 km/s), the two galaxies will experience a close encounter at a time $\Delta t = 6 \times 10^9$ yr in the future (i.e., at η = 10.2). Whether this encounter will lead to a collision/merger or simply a distant passage depends on the tangential velocity component, which is not well determined. The models of Peebles (1994) suggest that the distance of closest approach will lie in the range 20 – 416 kpc, with more models predicting values near the upper end of this range. Thus, more work is necessary to determine whether or not the Milky Way is destined to collide with M31 in the relatively near future.

However, even if our galaxy does not collide with M31 on the first pass, the two galaxies are clearly a bound binary pair. The orbits of binary galaxy pairs decay relatively

rapidly through dynamical friction (e.g., Binney & Tremaine, 1987; Weinberg, 1989). Thus, even if a collision does not occur on the first passing, M31 and the Milky Way will not survive very long as individual spiral galaxies. On a time scale of approximately $\eta = 11 - 12$, the entire local group will coalesce into one large stellar system.

B Gravitational radiation and the decay of orbits

Gravitational radiation acts in the opposite direction: it causes orbits to lose energy and decay so that the stars move inward. We first consider the case of a galaxy and its constituent stars. As a given star moves through the potential well of a galaxy, its orbit decays through gravitational radiation (e.g., Misner, Thorne, & Wheeler, 1973; Weinberg, 1972). The rate of energy loss is proportional to the square of the quadrapole moment of the whole system (see also Ohanian & Ruffini, 1994). For the case in which the galaxy has a large scale quadrapole moment (e.g., a bar), the rate of energy loss from gravitational radiation can be written in the simple form

$$\frac{\dot{E}}{E} = \left(\frac{\upsilon}{c}\right)^5 \tau^{-1}, \tag{3.4}$$

where $\tau = 2\pi R/\upsilon$ is the orbit time. For a galaxy, the rotation curve is almost flat with a nearly constant velocity $\upsilon \sim 200$ km/s. The time scale τ_{GR} for gravitational radiation is thus given by

$$\tau_{GR} = \frac{2\pi R}{\pi}\left(\frac{\upsilon}{c}\right)^{-5} \approx 10^{24}\,\mathrm{yr}\left(\frac{R}{R_0}\right), \tag{3.5}$$

where $R_0 = 10$ kpc is a reference length scale for the galaxy. We thus obtain the estimate

$$\eta_{GR} = 24 + \log_{10}\left[R/10\mathrm{kpc}\right]. \tag{3.6}$$

This time scale corresponds to $\sim 10^{16}$ orbits around the galactic center. Notice that if the stars are radiating incoherently in a nearly smooth potential, the time scale becomes longer by a factor of M_{gal}/M_*, where M_* is the mass of the star and M_{gal} is effective galactic mass. Notice also that gravitational orbital decay takes substantially longer than stellar evaporation from the galaxy (see the previous section). Thus, the evolution of the galaxy will be dominated by the collisionless process, and hence the majority of stellar remnants will be ejected into intergalactic space rather than winding up in the galactic core (see also Islam, 1977; Dyson, 1979; Rees, 1984).

Gravitational radiation also causes the orbits of binary stars to lose energy and decay. Of particular importance is the decay of binary brown dwarf stars. The eventual coalescence of these systems can lead to the formation of a new hydrogen burning star, provided that the mass of the entire system is larger than the hydrogen burning limit $M_H \sim 0.08 M_e$. The time scale τ_{OD} for orbital decay can be written

$$\tau_{OD} = \frac{\pi}{2}\frac{c^5 R_0^4}{G^3 M_\star^3}, \tag{3.7}$$

where $M_\star$ is the mass of the stars and R_0 is the initial orbital separation. Inserting numerical values and writing the result in terms of cosmological decades, we obtain the result

$$\eta_{OD} = 19.4 + 4\log_{10}\left[R_0 / (1\mathrm{AU})\right] - 3\log_{10}\left[M_\star / (1M_\odot)\right]. \tag{3.8}$$

This result also applies to planetary orbits (see §III.F below).

C Star formation through brown dwarf collisions

Once all of the interstellar material has been used up, one viable way to produce additional stars is through the collisions of brown dwarfs. These objects have masses too small for ordinary hydrogen burning to take place and hence their supply of nuclear fuel will remain essentially untapped. Collisions between these substellar objects can produce stellar objects with masses greater than the hydrogen burning limit, i.e., stars of low mass. We note that the search for brown dwarfs has been the focus of much observational work (see, e.g., Tinney, 1995) and the existence of these objects is now on firm ground (e.g., Golimowski et al., 1995; Oppenheimer et al., 1995).

1 Collision time scales

After conventional star formation in the galaxy has ceased, the total number of brown dwarfs in the galaxy will be N_0. Although the value of N_0 is uncertain and is currently the subject of much current research (e.g., see Alcock et al., 1993; Aubourg et al., 1993; Tinney, 1995), we expect that N_0 is roughly comparable to the number of ordinary stars in the galaxy today, $N_0 \sim 10^{11}$ (see §II. C). The rate Γ at which these brown dwarfs collide is given by

$$\Gamma = \frac{N\sigma\upsilon}{V} = -\frac{1}{N}\frac{dN}{dt}, \tag{3.9}$$

where N is the number of brown dwarfs in a galaxy with volume V, σ is the collision cross section (see below), and υ is the typical relative velocity. This equation can be integrated to obtain

$$N(t) = \frac{N_0}{1 + t/\tau_C}, \tag{3.10}$$

where τ_C is the characteristic time scale

$$\tau_C = \Gamma^{-1} \sim 10^{22}\,\mathrm{yr}, \tag{3.11}$$

or, equivalently,

$$\eta_C = 22 + \log_{10}\left[V / \left(20\text{kpc}\right)^3 \right] - \log_{10}\left[\upsilon / 200\text{km/s}\right]. \tag{3.12}$$

To obtain this numerical value for the time scale, we have assumed that the collision cross section is given by the geometrical cross section of the brown dwarfs; this assumption is justified below. We have also the used numerical values $V \sim (20\text{kpc})^3$ and $\upsilon \sim 200$ km/s which are characteristic of the galactic halo.

The estimate of collision rates given here is somewhat conservative. Nearby stellar encounters can lead to the formation of binaries through tidal excitation of modes on the stars (see Press & Teukolsky, 1977; Lee & Ostriker, 1986). These binaries can eventually decay and thereby lead to additional stellar collisions.

The time scale [3.12] is the time required for the halo population of brown dwarfs to change. Notice that this time scale is larger than the evaporation time scale calculated in §III.A. This ordering makes sense because distant encounters (which lead to evaporation) must be much more frequent than true collisions. For $\eta < \eta_C$, the collision rate of brown dwarfs for the entire galaxy is given by $\Gamma_{tot} = N/\tau_C \sim 10^{-11}$ yr^{-1}. The typical outcome of a brown dwarf collision will be the production of a stellar object with mass $M_* \sim 0.1 M_e$, large enough to burn hydrogen. The stellar (main-sequence) lifetime of such a star is roughly 2.5×10^{13} yr. This stellar evolutionary time scale is longer than the time scale on which stars are forming. As a result, the galaxy will produce many stars through this process and will contain ~ 100 hydrogen burning stars for cosmological decades $\eta > 14$.

Notice that the time scale for producing stars through brown dwarf collisions is generally much shorter than the orbit decay time for brown dwarf binaries. For orbital decay, equation [3.8] implies that $\eta \sim 22.5 + 4\log_{10}(R/1\text{AU})$. Thus, brown dwarf collisions provide the dominant mechanism for continued star formation while the galaxy remains intact.

2 Collision cross sections

To complete this argument, we must estimate the cross section for colliding brown dwarfs. Consider two brown dwarfs with a relative velocity υ_{rel}. For simplicity, we consider the case of equal mass brown dwarfs with mass m. The orbital angular momentum of the system is given by

$$J = m\upsilon_{rel} b, \tag{3.13}$$

where b is the impact parameter. When the two dwarfs collide and form a composite star of mass ~$2m$, the angular momentum can be written

$$I\Omega = f\left(2m\right)R^2\Omega, \tag{3.14}$$

where R is the stellar radius, Ω is the rotation rate, and f is a numerical constant of order unity which depends on the internal structure of the star. We next invoke the

constraint that the rotation rate of the final state must be less than the break-up speed, i.e.,

$$\Omega^2 R^2 < \frac{G(2m)}{R}. \tag{3.15}$$

Combining the above results, we obtain a bound on the impact parameter b that can lead to a bound final system. We thus obtain

$$b^2 < \frac{8 f^2 GmR}{\upsilon_{rel}^2}, \tag{3.16}$$

which can be used to estimate the cross section,

$$\sigma \approx \pi b^2 = \frac{8\pi f^2 GmR}{\upsilon_{rel}^2}. \tag{3.17}$$

Using typical numerical values, we find that $b \sim R \sim 10^{10}$ cm, which is roughly comparable to the radius of the brown dwarf (e.g., Burrows et al., 1993).

3 Numerical simulations and other results

In order to illustrate the viability of this collision process, we have done a set of numerical simulations using smooth particle hydrodynamics (SPH). We find that collisions between substellar objects can indeed form final products with masses greater than the minimum mass required to burn hydrogen. Examples of such collisions are shown in Figure 5.3. In these simulations, density structures from theoretical brown dwarf models (Laughlin & Bodenheimer, 1993) are delivered into impact with relative velocity 200 km/s. The hydrodynamic evolutionary sequences shown are adiabatic. One expects that the emergent stellar mass object will contract toward the main sequence on a Kelvin-Helmholtz time scale and then initiate hydrogen burning.

Finally, we note that white dwarfs will also collide in the galactic halo. As outlined in §II.E, we expect roughly comparable numbers of white dwarfs and brown dwarfs at the end of the stelliferous era. Although the white dwarfs are actually smaller in radial size, they are more massive and hence have a larger gravitational enhancement to their interaction cross section. As a result, the net cross section and hence the net interaction rate of white dwarfs should be roughly comparable to that of brown dwarfs (§III.C.1). When white dwarfs collide with each other, several different final states are possible, as we discuss below.

If the two white dwarfs are sufficiently massive, it is possible that the collision product will have a final mass which exceeds the Chandrasekhar limit ($M_{Ch} \approx 1.4 M_e$) and hence can explode in a supernova. Using the final mass function (see §II.E and Figure 5.2), we estimate that roughly one third of the white dwarfs will have masses greater than $0.7 M_e$ and hence only about one tenth of the collisions can possibly result in an object exceeding the Chandrasekhar mass limit. The supernova rate from these

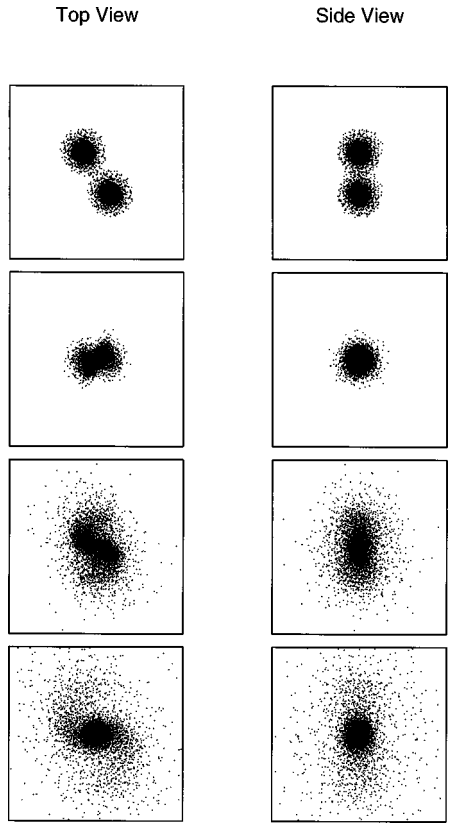

Figure 5.3 Numerical simulation of a collision between two brown dwarfs. The two initial objects have masses less than that required for hydrogen burning; the final product of the collision is a true star and is capable of sustained hydrogen fusion. The two stars collide with a relative velocity of 200 km/s and an impact parameter of ~ 1 stellar radius. The top series of panels shows the collision from a side view; the bottom series of panels shows the top view.

collisions can thus be as large as $\Gamma_{SN} \sim 10^{-12}$ yr^{-1}, although it will be somewhat smaller in practice due to inefficiencies.

The most common type of collision is between two low-mass white dwarfs – the final mass function peaks at the mass scale $M_* \approx 0.13 M_e$. These low-mass objects will have an almost pure helium composition. If the final product of the collision has a mass larger than the minimum mass required for helium burning ($M_{He} \approx 0.3 M_e$), then the product star could land on the helium main-sequence (see, e.g., Kippenhahn & Weigert, 1990). In order for the star to burn helium, the collision must be sufficiently energetic to impart enough thermal energy into the star; otherwise, the star will become just another helium white dwarf. Another possibility exists for collisions between white dwarfs of slightly larger masses. If the product of the collision has a mass smaller than

the Chandrasekhar mass and larger than the minimum mass to burn carbon ($0.9M_e \leq M_* \leq 1.4M_e$), the product star could land on the carbon main sequence. Thus, this mode of late time star formation can lead to an interesting variety of stellar objects.

D The black hole accretion time

Large black holes tend to accrete stars and gas and thereby increase their mass. The black hole accretion time is the characteristic time scale for a black hole in the center of a galaxy to swallow the rest of the galaxy. If we consider collisions of the black hole with stars, and ignore the other processes discussed above (gravitational radiation and stellar evaporation), the time for the black hole to absorb the stars in the galaxy is given by

$$\tau = \frac{V}{\sigma \upsilon}, \tag{3.18}$$

where $V = R^3$ is the volume of the galaxy, υ is the typical speed of objects in the galaxy ($\upsilon \sim 200$ km/s), and σ is the effective cross section of the black hole. As a starting point, we write the cross section in the form

$$\sigma = \Lambda \pi R_S^2, \tag{3.19}$$

where Λ is a dimensionless enhancement factor due to gravitational focusing, and R_S is the Schwarzschild radius R_S given by

$$R_S = GM / c^2. \tag{3.20}$$

We thus obtain the time scale

$$\tau = 10^{30}\,\mathrm{yr}\left[M / 10^6 M_\odot\right]^{-2}\left[R / 10\mathrm{kpc}\right]^3 \Lambda^{-1}, \tag{3.21a}$$

or, equivalently,

$$\eta_{\mathrm{accrete}} = 30 - 2\log_{10}\left[M / 10^6 M_\odot\right] + 3\log_{10}\left[R / 10\mathrm{kpc}\right] - \log_{10}\left[\Lambda\right]. \tag{3.21b}$$

The time scale $\eta_{accrete} \sim 30$ is much longer than the time scale for both stellar evaporation and gravitational radiation (see also the following section). As a consequence, at these late times, all the stars in a galaxy will either have evaporated into intergalactic space or will have fallen into the central black hole via gravitational radiation decay of their orbits. Of course, as the black hole mass grows, the accretion time scale decreases. Very roughly, we expect ~ 1 – 10% of the stars to fall to the central black hole and the remainder to be evaporated; the final mass of the central black hole will thus be $M_{BH} \sim 10^9 - 10^{10}\, M_e$.

One can also consider this process on the size scale of superclusters. When $\eta \sim 30$, supercluster-sized cosmological density perturbations of length R will have long since evolved to nonlinearity, and will be fully decoupled from the Hubble flow. One can imagine an ensemble of $\sim 10^9 \sim 10^{10}\, M_e$ black holes which have descended from dead galaxies and are now roaming freely and hoovering up an occasional remaining star in the volume R^3. The characteristic time scale for this process is

$$\eta_{\rm accrete} = 33 - 2\log_{10}\left[M / 10^9\, M_\odot \right] + 3\log_{10}\left[R / 10{\rm Mpc} \right]. \tag{3.22}$$

As for the case of the galaxy, however, this straightforward scenario is compromised by additional effects. Gravitational radiation will continuously cause the orbits of the black holes to decay, and some of them may eventually merge. Stellar encounters with both other stars and with the black holes will lead to stellar evaporation from the supercluster sized system. Over the long term, one expects that the supercluster will consist of a very large central black hole with a majority of the stars and many of the original $\sim 10^9 - 10^{10}\, M_e$ galactic black holes escaping to large distances. In other words, the supercluster-sized system will behave somewhat analogously to the galaxy, except that it will contain a larger size scale, a longer time scale, and two widely disparate mass scales (namely, a stellar mass scale $M_\star \sim 1 M_e$, and a black hole mass scale $M_{BH} \sim 10^9 - 10^{10}\, M_e$). Equipartition effects between the two mass scales will come into play, and will drive the galactic black holes toward the center while preferentially ejecting the stellar remnants. In principle, this hierarchy can extend up to larger and larger perturbation length scales, although the relevant time scales and detailed dynamics become more uncertain as one proceeds with the extrapolation.

E Annihilation and capture of halo dark matter

Galactic halos consist largely of dark matter, much of which may reside in non-baryonic form. Although the nature and composition of this dark matter remains an important open question, one of the leading candidates is Weakly Interacting Massive Particles, usually denoted as WIMPs. These particles are expected to have masses in the range $M_W = 10 - 100$ GeV and interact through the weak force and gravity only (cf. the reviews of Diehl et al., 1995; Jungman, Kamionkowski & Griest, 1996; see also the recent proposal of Kane & Wells, 1996). Many authors have studied the signatures of WIMP annihilation, usually with the hope of finding a detectable signal. One can apply the results of these studies to estimate the time scale for the depletion of WIMPs from a galactic halo.

We first consider the case of direct particle-particle annihilation. Following usual conventions, the rate Γ_W for WIMP annihilation in the halo can be written in the form

$$\Gamma_W = n_W \langle \sigma \upsilon \rangle, \tag{3.23}$$

where n_W is the number density of WIMPs in the halo and $\langle \sigma v \rangle$ is the average value of the annihilation cross section times velocity. If WIMPs make up a substantial mass

fraction of the galactic halo, their number density is expected to be roughly $n_W \sim 1$ cm^{-3}. The typical velocity of particles in the galactic halo is ~200 km/s. Using the most naive dimensional argument, we can estimate the interaction cross section as

$$\sigma \sim M_W^2 G_F^2 \sim 5\times 10^{-38}\,\mathrm{cm}^2 \left[\frac{M_W}{1\mathrm{GeV}}\right]^2, \tag{3.24}$$

where M_W is the mass of the particle and G_F is the Fermi constant. The true cross section has additional factors which take into account spin dependences, mixing angles, and other model dependent quantities (see Diehl et al., 1995; Jungman et al., 1996); the form [3.24] is thus highly approximate, but adequate for our purposes. We also note that the relic abundance of dark matter particles is determined by the interaction cross section; in order for the abundance to be cosmologically significant, the interaction cross section must be of order $\sigma \sim 10^{-37}$ cm^2 (see Kolb & Turner, 1990).

Putting all of the above results together, we can estimate the time scale τ_W for the population of WIMPs to change,

$$\tau_W = \Gamma^{-1} = \frac{1}{n_W \langle \sigma \rangle} \sim 3\times 10^{22}\,\mathrm{yr}. \tag{3.25}$$

Thus, in terms of cosmological decades, we obtain the annihilation time scale in the form

$$\eta_W = 22.5 - \log_{10}\left[\frac{\langle \sigma \upsilon \rangle}{10^{-30}\,\mathrm{cm}^3 s^{-1}}\right] - \log_{10}\left[\frac{n_W}{1\mathrm{cm}^{-3}}\right]. \tag{3.26}$$

It takes a relatively long time for WIMPs to annihilate via direct collisions. In particular, the annihilation time scale is much longer than the stellar evaporation time scale (§III.A).

Another important related effect is the capture of WIMPs by astrophysical objects. The process of WIMP capture has been studied for both the Sun (Press & Spergel, 1985; Faulkner & Gilliland, 1985) and the Earth (Freese, 1986) as a means of helping to detect the dark matter in the halo (see also Krauss, Srednicki, & Wilczek, 1986; Gould, 1987, 1991). Although WIMP capture by the Sun and the Earth can be important for dark matter detection, the lifetimes of both (main sequence) stars and planets are generally too small for WIMP capture to significantly affect the total population of particles in the galactic halo. On the other hand, stellar remnants, in particular white dwarfs, can be sufficiently long lived to have important effects.

In astrophysical objects, WIMPs are captured by scattering off of nuclei. When the scattering event leads to a final velocity of the WIMP that is less than the escape speed of the object, then the WIMP has been successfully captured. For the case of white dwarfs, we can make the following simple estimate of the capture process. The mean free path of a WIMP in matter with white dwarf densities is generally less than the radius of the star. In addition, the escape speed from a white dwarf is large, roughly

~ 3000 km/s, which is much larger than the velocity dispersion of WIMPs in the halo. As a result, to first approximation, most WIMPs that pass through a white dwarf will be captured. The WIMP capture rate Γ_{W*} by a white dwarf is thus given by

$$\Gamma_{W.} = n_W \sigma_W D^{\upsilon} rel, \tag{3.27}$$

where $\sigma_{WD} \sim 10^{18}$ cm^2 is the cross sectional area of the white dwarf and $\upsilon_{rel} \sim 200$ km/s is the relative velocity. The capture rate is thus

$$\Gamma_{W.} \sim 10^{25} s^{-1}. \tag{3.28}$$

With this capture rate, a white dwarf star can consume its weight in WIMPs on a time scale of $\sim 10^{24}$ yr. The total mass in WIMPs in the halo is expected to be a factor of 1–10 times the mass of stars, which will be mostly in the form of white dwarfs at these late times (§II.E). As a result, the time scale for white dwarfs to deplete the entire halo population of WIMPs via capture is roughly given by

$$\tau \sim 10^{25} \text{yr} \quad \text{or} \quad \eta \sim 25. \tag{3.29}$$

The actual time scales will depend on the fraction of the galactic halo in non-baryonic form and on the properties (e.g., mass) of the particles; these quantities remain unknown at this time.

The annihilation of halo WIMPs has important consequences for both the galaxy itself and for the white dwarfs. Basically, the galaxy as a whole loses mass while the white dwarfs are kept hotter than they would be otherwise. The population of captured WIMPs inside the star will build up to a critical density at which the WIMP annihilation rate is in equilibrium with the WIMP capture rate (see, e.g., Jungman et al., 1996). Furthermore, most of the annihilation products will be absorbed by the star, and the energy is eventually radiated away (ultimately in photons). The net result of this process (along with direct annihilation) is thus to radiate away the mass of the galactic halo on the time scales given by equations [3.26] and [3.29]. This process competes with the evaporation of stars through dynamical relaxation (§III.A) and the decay of stellar orbits through gravitational radiation (§III.B).

Since the time scale for WIMP evaporation is much longer than the dynamical time scale, the galaxy will adiabatically expand as the halo radiates away. In the outer galaxy, the dark matter in the halo dominates the gravitational potential well and hence the stars in the outer galaxy will become unbound as the halo mass is radiated away. Since WIMPs do not dominate the potential inside the solar circle, the corresponding effects on the inner galaxy are relatively weak.

The white dwarf stars themselves will be kept hot by this WIMP capture process with a roughly constant luminosity given by

$$L_{WD} = F m_W \Gamma_{W.} = F m_W n_W \sigma_W D^{\upsilon} rel \sim 4 \times 10^{-12} L_{\odot}, \tag{3.30}$$

where F is an efficiency factor (expected to be of order unity) which takes into account the loss of energy from the star in the form of neutrinos. With this luminosity, the white dwarf has a surface temperature $T_* \approx 63$ K, where we have assumed a typical white dwarf mass $M_* = 0.5M_e$. As a reference point, we note that an entire galaxy of such stars has a total luminosity comparable to that of the sun, $L_{gal} \sim 1L_e$. However, most of the radiation will be emitted at infrared wavelengths, $\lambda \sim 50\mu$m.

For completeness, we note that axions provide another viable candidate for the dark matter in the galactic halo (see Chapter 10 of Kolb & Turner, 1990). These particles arise from solutions to the strong CP problem in quantum chromodynamics (see, e.g., Peccei & Quinn, 1977ab; Weinberg, 1978; Wilczek, 1978). The coupling of the axion to the photon allows the axion to decay to a pair of photons with a lifetime τ_a given by

$$\tau_a \approx 2\times 10^{17}\,\mathrm{yr}\left(m_a / 1\mathrm{eV}\right)^{-5}, \tag{3.31}$$

where m_a is the mass of the axion; we have assumed representative values for the remaining particle physics parameters. Relic axions with sufficient numbers to contribute to the dark matter budget of the universe have masses in the range 10^{-6} eV $< m_a < 10^{-3}$ eV, where the value depends on the production mechanism. Using these mass values, we obtain an allowed range of axion decay time scales,

$$32 \le \eta_a \le 47. \tag{3.32}$$

F The fate of planets during galactic death

Planets can be loosely defined as objects that are small enough (in mass) to be supported by ordinary Coulomb forces rather than by degeneracy pressure. Over the long term, planets suffer from several deleterious processes. They can be vaporized by their evolving parent stars, and their orbits can either decay or be disrupted. Barring these more imminent catastrophes, planets will evaporate as their protons decay (see §IV.H).

The theory of general relativity indicates that planetary orbits slowly decay via emission of gravitational radiation (see §III.B). To fix ideas, consider a planet orbiting a star of mass M_* at an initial orbital radius R. Gravitational radiation drives orbital decay on a time scale given by

$$\tau = \frac{2\pi R}{\upsilon}\left(\frac{\upsilon}{c}\right)^{-5} = 2.6\times 10^{19}\,\mathrm{yr}\left(\frac{R}{1\mathrm{AU}}\right)^{4}\left(\frac{M_*}{1M_\odot}\right)^{-3}, \tag{3.33}$$

or, in terms of cosmological decades,

$$\eta = 19.4 + 4\log_{10}\left[R / 1\mathrm{AU}\right] - 3\log_{10}\left[M_* / 1M_\odot\right]. \tag{3.34}$$

In the interim, planets can be dislodged from their parent stars during encounters and collisions with interloping stars. The time scale for these dislocations is given by

the time interval required to produce a sufficiently close encounter with another star. Very roughly, if a perturbing star intrudes within a given planet's orbit, then the planet is likely to be entirely ejected from the system. This time scale is given by

$$\tau = \frac{1}{n\sigma\upsilon}, \tag{3.35}$$

where n is the number density of stars (~ 0.1 pc^{-3} in our galaxy today), υ is the relative velocity (~100 km/s), and where the cross section σ is determined by the orbital radius of the planet ($\sigma \approx \pi R^2$). Inserting these values, one finds

$$\tau = 1.3\times10^{15}\,\mathrm{yr}\left(\frac{R}{1\mathrm{AU}}\right)^{-2}, \tag{3.36}$$

$$\eta = 15.1 - 2\log_{10}\left[R/1\mathrm{AU}\right], \tag{3.37}$$

where R is the radius of the planetary orbit.

Comparing equation [3.33] with equation [3.36], we find that the time scale for gravitational radiation is equal to that of stellar encounters for planetary orbits of radius R = 0.2 AU, which is about half the radius of the orbit of Mercury in our own solar system. One might guess then, that very close planets, such as the recently discovered companion to 51 Pegasus (Mayor & Queloz, 1995; Marcy, Butler, & Williams, 1996), will eventually merge with their parent stars as a result of radiative orbital decay, while planets with larger initial orbits (e.g., the giant planets in our solar system) will be stripped away from their parent stars as a consequence of stellar encounters. However, since the time scale for stellar evolution ($\eta_* < 14$) is much shorter than the time scale for orbital decay, close-in planets around solar-type stars will be destroyed during the red giant phases long before their orbits are constricted by general relativity. Only the inner planets of low-mass M dwarfs (which experience no giant phases) will find their fate sealed by gravitational radiation.

IV Long-term fate of degenerate stellar objects

Brown dwarfs, white dwarfs, neutron stars, and black holes have lifetimes which are not only much longer than the current age of the universe (η = 10), but also greatly exceed the expected lifetime of the galaxy (η = 20 – 25). Due to a general lack of urgency, the ultimate fate of these objects has not yet been extensively considered. Nevertheless, these objects will not live forever. If the proton is unstable, then proton decay will drive the long-term evolution of degenerate stellar objects. Black holes are essentially unaffected by proton decay, but they gradually dissipate via the emission of Hawking radiation. Both proton decay and Hawking radiation yield many interesting astrophysical consequences. In the following discussion, we work out the details of these processes (see also Dicus et al., 1982; Feinberg, 1981).

A Proton decay

In Grand Unified Theories (GUTs), the proton is unstable and has a finite, albeit quite long, lifetime. For example, the proton can decay through the process

$$p \rightarrow e^{+} + \pi^{0}\,, \tag{4.1}$$

and the Feynman diagram for this decay process is shown in Figure 5.4. Many different additional decay channels are possible and the details ultimately depend on the particular theory (e.g., see the reviews of Langacker, 1981; Perkins, 1984). In particular, we note that many other decay products are possible, including neutrinos. If protons are unstable, then neutrons will also be unstable over a commensurate time scale. Free neutrons are of course unstable to β-decay on a very short time scale (~ 10 minutes); however, bound neutrons will be unstable through processes analogous to the proton decay modes (e.g., see Figure 5.4). In the present context, the protons and neutrons of interest are bound in "heavy" nuclei (mostly carbon and helium) within white dwarfs.

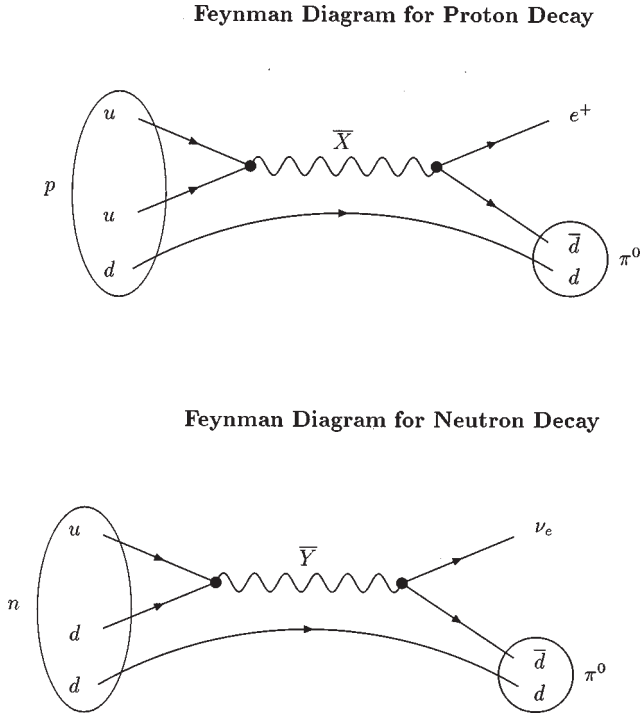

Figure 5.4 Representative Feynman diagrams for proton decay (top diagram) and neutron decay (bottom diagram) shown in terms of the constituent quarks (u, d, $\bar{d}$). These processes are the form expected for the simplest Grand Unified Theories. The particles X and Y are the intermediate vector bosons which mediate the baryon number violating process and are expected to have masses comparable to the GUT scale ~ 10^{16} GeV.

For the simplest class of GUTs, as illustrated by the decay modes shown in Figure 5.4, the rate of nucleon decay Γ_P is roughly given by

$$\Gamma_P = \alpha_5^2 \frac{m_P^5}{M_X^4}, \tag{4.2}$$

where m_P is the proton mass and α_5 is a dimensionless coupling parameter (see, e.g., Langacker, 1981; Perkins, 1984; Kane, 1993). The mass scale M_X is the mass of the particle which mediates the baryon number violating process. The decay rate should also include an extra numerical factor which takes into account the probability that the interacting quarks (which participate in the decay) are in the same place at the same time; this numerical factor is less than unity so that the proton lifetime is larger by a corresponding factor. To a first approximation, the time scale for proton decay is thus given by

$$\tau_P \approx 10^{37}\ \mathrm{yr} \left[\frac{M_X}{10^{16}\mathrm{GeV}} \right]^4, \tag{4.3}$$

where we have taken into account the aforementioned numerical probability factor. The corresponding cosmological time scale is

$$\eta p = 37 + 4 \log_{10} \left[M_X / 10^{16}\,\mathrm{GeV} \right]. \tag{4.4}$$

Notice that this time scale has a very sensitive dependence on the mass scale M_X of the mediating boson.

We want to find the allowed range for the proton lifetime. This time scale is constrained from below by current experimental limits on the lifetime of the proton (e.g., Perkins, 1984). The proton lifetime must be greater than $\eta \sim 32$ (10^{32} yr), where the exact limit depends on the particular mode of proton decay (Particle Data Group, 1994). Finding an upper bound is more difficult. If we restrict our attention to the class of proton decay processes for which equation [4.4] is valid, then we must find an upper bound on the mass scale M_X. Following cosmological tradition, we expect the scale M_X to be smaller than the Planck scale $M_{Pl} \approx 10^{19}$ GeV, which implies the following range for the proton lifetime,

$$32 < \eta p < 49\,. \tag{4.5}$$

The lower bound is set by experimental data; the upper bound is more suggestive than definitive (see also §IV.F).

We can find a more restrictive range for the proton lifetime for the special case in which the decay mode from some GUT is responsible for baryogenesis in the early universe. (Note that some baryon number violating process is necessary for baryogenesis to take place – see Sakharov, 1967). Let us suppose that the decay mode from some GUT is valid and that baryogenesis takes place at an energy scale in the

early universe $E_B \sim M_X$. This energy scale must be less than the energy scale E_I of the inflationary epoch (Guth, 1981). The inflationary energy scale is constrained to be less than $\sim 10^{-2}\, M_{PI}$ in order to avoid overproducing scalar density perturbations and gravitational radiation perturbations (Lyth, 1984; Hawking, 1985; Krauss & White, 1992; Adams & Freese, 1995). Combining these two constraints, we obtain the following suggestive range for the time scale for proton decay,

$$32 < \eta p < 41 \tag{4.6}$$

Although a range of nine orders of magnitude in the relevant time scale seems rather severe, the general tenor of the following discussion does not depend critically on the exact value. For the sake of definiteness, we adopt $\eta_P = 37$ as a representative time scale.

B White dwarfs powered by proton decay

On a sufficiently long time scale, the evolution of a white dwarf is driven by proton decay. When a proton decays inside a star, most of the primary decay products (e.g., pions and positrons) quickly interact and/or decay themselves to produce photons. For example, the neutral pion π^0 decays into a pair of photons with a lifetime of $\sim 10^{-16}$ sec; positrons, e^+, last only $\sim 10^{-15}$ sec before annihilating with an electron and producing gamma rays. Therefore, one common net result of proton decay in a star is the eventual production of four photons through the effective reaction

$$p + e^{-1} \rightarrow \gamma + \gamma + \gamma + \gamma, \tag{4.7}$$

where the typical energy of the photons is given by $E\gamma \sim m_P/4 \sim 235$ MeV. These photons have a relatively short mean free path within the star and will thermalize and diffuse outwards through a random walk process with a characteristic time scale of $\sim 10^5$ yr, much shorter than the evolutionary time scale of the system. Additionally, some fraction of the decay products are in the form of neutrinos, which immediately leave the system.

When proton decay is a white dwarf's primary energy source, the luminosity is

$$L_\star(t) = FN_0 E\Gamma_P e^{-\Gamma_P t} \approx FM(t)\Gamma_P, \tag{4.8}$$

where $N_0 \sim 10^{57}$ is the initial number of protons in the star, $E \sim 1$ GeV is the net energy produced per decay, and Γ_P is the decay rate. The factor F is an efficiency parameter which takes into account the fraction of energy lost in the form of neutrinos. Very roughly, we expect $\sim 1/3$ of the energy in the decay products to be in neutrinos and hence $F \sim 2/3$ (e.g., Dicus et al., 1982). The exact value of the fraction F depends on the branching ratios for a particular GUT and hence is model dependent. For a typical decay rate of $\Gamma_P \sim 10^{-37}\ \mathrm{yr}^{-1}$, the luminosity in solar units becomes

$$L_\star \sim 10^{-24} L_\odot . \tag{4.9}$$

It is perhaps more illuminating to express this stellar luminosity in ordinary terrestrial units. A white dwarf fueled by proton decay generates approximately 400 Watts, enough power to run a few light bulbs, or, alternately, about 1/2 horsepower. An entire galaxy of such stars has a total luminosity of $L_{gal} \sim 10^{-13} L_e$, which is much smaller than that of a single hydrogen burning star.

The total possible lifetime for a star powered by proton decay is given by

$$\tau = \frac{1}{\Gamma_P} \ln\left[N_0 N_{\min} \right], \tag{4.10}$$

where $N_0 \sim 10^{57}$ is the initial number of nucleons in the star and N_{min} is the minimum number of nucleons required to consider the object a star. If, for example, one takes the extreme case of $N_{min} = 1$, the time required for the star to completely disappear is $t \approx 130/\Gamma_P$; in general we obtain

$$\eta_* = \eta_P + \log_{10}\left[\ln\left(N_0 / N_{\min} \right) \right]. \tag{4.11}$$

As we show in §IV.D, the object ceases to be a star when $N_{min} \sim 10^{48}$ and hence $\eta_* \approx \eta_P + 1.3$.

During the proton decay phase, the stellar surface temperature is given by

$$T_*^4 = \frac{F N_0 E \Gamma_P}{4\pi\sigma_B R_*^2} \mathrm{e}^{-\Gamma_P t} \tag{4.12}$$

where we have assumed that the spectral energy distribution is simply a blackbody (σ_B is the Stefan-Boltzmann constant). For a $1 M_e$ star and the typical decay rate Γ_P, the effective stellar temperature is $T_* \sim 0.06$ K. This temperature will be enormously hotter than the temperature of the universe's background radiation at the cosmological decade $\eta = 37$.

As a white dwarf loses mass via proton decay, the star expands according to the usual mass/radius relation

$$R_* M_*^{1/3} = 0.114 \frac{h^2}{G m_e m_P^{5/3}} \left(Z / A \right)^{5/3} \tag{4.13}$$

where Z and A are the atomic number and atomic weight of the white dwarf material (e.g., Chandrasekhar, 1939; Shu, 1982; Shapiro & Teukolsky, 1983). For simplicity, we will take typical values and use $A = 2Z$. If we also rewrite the white dwarf mass/radius relation in terms of natural units, we obtain the relation

$$R_* = 1.42 \left(\frac{M_{\mathrm{Pl}}}{m_e} \right) \left(\frac{M_{\mathrm{Pl}}}{m_P} \right) \left(\frac{M_*}{m_P} \right)^{-1/3} m_P^{-1}. \tag{4.14}$$

While the white dwarf is in the proton decay phase of its evolution, the star follows a well defined track in the H-R diagram, i.e.,

$$L_\star = L_0 \left(T_\star / T_0 \right)^{12/5} . \tag{4.15}$$

or, in terms of numerical values,

$$L_\star = 10^{-24} L_\odot \left[\frac{T_\star}{0.06\,\mathrm{K}} \right]^{12/5} . \tag{4.16}$$

We note that the white dwarf mass/radius relation depends on the star's chemical composition, which changes as the nucleons decay (see the following section). This effect will cause the evolutionary tracks to depart slightly from the 12/5 power-law derived above. However, this modification is small and will not be considered here.

C Chemical evolution in white dwarfs

Over the duration of the proton decay phase, the chemical composition of a white dwarf is entirely altered. Several different effects contribute to the change in chemical composition. The nucleon decay process itself directly alters the types of nuclei in the star and drives the chemical composition toward nuclei of increasingly lower atomic numbers. However, pycnonuclear reactions can occur on the relevant (long) time scales and build nuclei back up to higher atomic numbers. In addition, spallation interactions remove protons and neutrons from nuclei; these free nucleons then interact with other nuclei and lead to further changes in composition.

In the absence of pycnonuclear reactions and spallation, the chemical evolution of a white dwarf is a simple cascade toward lower atomic numbers. As protons and neutrons decay, the remaining nuclei become correspondingly smaller. Some of the nuclear products are radioactive and will subsequently decay. Given the long time scale for proton decay, these radioactive nuclei are extremely short-lived. As a result, only the stable isotopes remain. At relatively late times, when the total mass of the star has decreased by a substantial factor (roughly a factor of ten as we show below), almost all of the nuclei left in the star will be in the form of hydrogen.

At high densities and low temperatures, nuclear reactions can still take place, although at a slow rate. The quantum mechanical zero point energy of the nuclei allows them to overcome the Coulomb repulsion and fuse. In natural units, the nuclear reaction rate can be written in the form

$$W = 4 \left(\frac{2}{\pi^3} \right)^{1/2} S\left(Z^2 \alpha \mu \right)^{3/4} R_0^{-5/4} \exp\left[-4Z \left(\alpha \mu R_0 \right)^{1/2} \right], \tag{4.17}$$

where μ is the reduced mass of the nucleus, R_0 is the average spacing between nuclei, and α is the fine structure constant (see Shapiro & Teukolsky, 1983). A slightly different form for this reaction rate can be derived by including anisotropic and electron screening effects (Salpeter & Van Horn, 1969), but the basic form is similar. The parameter $S(E)$ is a slowly varying function of energy which takes into account the

probability of two nuclei interacting given that tunneling has occurred. Specifically, the parameter S is related to the cross section $\sigma(E)$ through the relation

$$\sigma(E) = \frac{S(E)}{E} T \,, \tag{4.18}$$

where T is the tunneling transition probability. The parameter S can be determined either from direct experiments or from theoretical calculations (see Shapiro & Teukolsky, 1983; Bahcall, 1989).

In order to evaluate the time scale for pycnonuclear reactions to occur, one needs to determine the spacing R_0 of the nuclei, or, equivalently, the number density of particles. Using the white dwarf mass/radius relation, we obtain the result

$$\mu R_0 = 2.29 A \left(\frac{M_{\rm Pl}}{m_e} \right) \left(\frac{M_{\rm Pl}}{m_P} \right) \left(\frac{M_*}{m_P} \right)^{-2/3} \approx 4060 A m_*^{-2/3} \,, \tag{4.19}$$

where A is average the atomic weight of the nuclei and where we have defined $m_* \equiv M_*/M_e$.

We can now obtain a rough estimate for the efficiency of pycnonuclear reactions building larger nuclei within white dwarfs. As a reference point, we note that for a density of $\rho \sim 10^6$ g cm^{-3}, the time scale for hydrogen to fuse into helium is $\sim 10^5$ yr (e.g., Shapiro & Teukolsky, 1983; Salpeter & van Horn, 1969), which is much shorter than the proton decay time scale. However, the form of equation [4.17] shows that the rate of nuclear reactions becomes highly suppressed as the reacting nuclei become larger. The exponential suppression factor roughly has the form $\sim \exp[-\beta Z A^{1/2}]$, where the numerical factor $\beta \approx 22$. Thus, as the quantity $ZA^{1/2}$ increases, the rate of nuclear reactions decreases exponentially. For example, if $Z = 6$ and $A = 12$ (for carbon), this exponential term is a factor of $\sim 10^{-190}$ smaller than that for hydrogen. Because of this large exponential suppression, fusion reactions will generally not proceed beyond helium during the late time chemical evolution considered here. Thus, the net effect of pycnonuclear reactions is to maintain the decaying dwarf with a predominantly helium composition down to a lower mass scale.

Spallation is another important process that affects the chemical evolution of white dwarf stars during the epoch of proton decay. The high energy photons produced through proton decay can interact with nuclei in the star. The most common result of such an interaction is the emission of a single free neutron, but charged particles (protons), additional neutrons, and gamma rays can also result (e.g., Hubbell, Gimm, & Overbo, 1980). The free neutrons will be promptly captured by other nuclei in a type of late time *s*-process (the *r*-process is of course dramatically irrelevant). The free protons can produce heavier nuclei through pycnonuclear reactions, as described above. Both of these mechanisms thus allow heavier elements to build up in the star, albeit at a very slow rate and a very low abundance. Thus, the process of spallation initially produces free neutrons and protons; but these nucleons are incorporated into other nuclei. As a result, the net effect of spallation is to remove nucleons from some nuclei and then give them

back to other nuclei within the star. The result of this redistribution process is to widen the distribution of the atomic numbers (and atomic weights) for the nuclei in the star.

In order to assess the importance of spallation processes, we must consider the interaction cross section. To leading order, the cross section for nuclear absorption of photons is a single "giant resonance" with a peak at about 24 MeV for light nuclei and a width in the range $\Gamma = 3 - 9$ MeV. The relative magnitude of this resonance feature is ~20 mb (see, e.g., Hubbell, Gimm, & Overbo, 1980; Brune & Schmidt, 1974), roughly a factor of 30 smaller than the total interaction cross section (which is dominated by scattering and pair production). For each proton decay event, ~940 MeV of matter is converted into photons, with some neutrino losses. When these photons cascade downward in energy through the resonance regime (at ~24 MeV), there will be 20 – 40 photons and about one in 30 will produce a spallation event. Hence, on average, each proton decay event leads to approximately one spallation event.

Spallation products allow the interesting possibility that a CNO cycle can be set up within the star. The time scale for pycnonuclear reactions between protons (produced by spallation) and carbon nuclei is short compared to the proton decay time scale. The time scale for pycnonuclear reactions between protons and nitrogen nuclei is comparable to the proton decay time scale. Thus, in principle, the white dwarf can set up a CNO cycle analogous to that operating in upper-main-sequence stars (see Clayton, 1983; Kippenhahn & Weigert, 1990; Shu, 1982). The energy produced by this cycle will be small compared to that produced by proton decay and hence this process does not actually affect the luminosity of the star. However, this cycle will affect the chemical composition and evolution of the star. As usual, the net effect of the CNO cycle is to build four free protons into a helium nucleus and to maintain an equilibrium abundance of the intermediate nitrogen and oxygen nuclei.

In order to obtain some understanding of the chemical evolution of white dwarfs, we have performed a simple numerical simulation of the process. Figure 5.5 shows the results of this calculation for a 1 M_e white dwarf with an initial chemical composition of pure carbon ^{12}C. The simulation assumes that radioactive isotopes decay immediately as they are formed through the preferred decay modes. For each proton decay event, a spallation event also occurs (see above) and leads to the removal of a nucleon from a random nucleus; the spallation products are then assumed to fuse immediately and randomly with other nuclei through the *s*-process and pycnonuclear reactions. The spallation process builds up a small abundance of nuclei heavier than the original ^{12}C, particularly ^{13}C which has a substantial mass fraction at "early" times. The white dwarf evolves through successive phases in which smaller and smaller nuclei are the dominant elements by mass fraction. The star never builds up a significant lithium fraction due to the immediate fission of newly formed ^{8}Be into α particles. The star has a broad phase during which ^{4}He dominates the composition. When the white dwarf has lost about 60% of its original mass, the hydrogen mass fraction begins to predominate.

D Final phases of white dwarf evolution

In the final phases in the life of a white dwarf, the star has lost most of its mass through proton decay. When the mass of the star becomes sufficiently small, two important

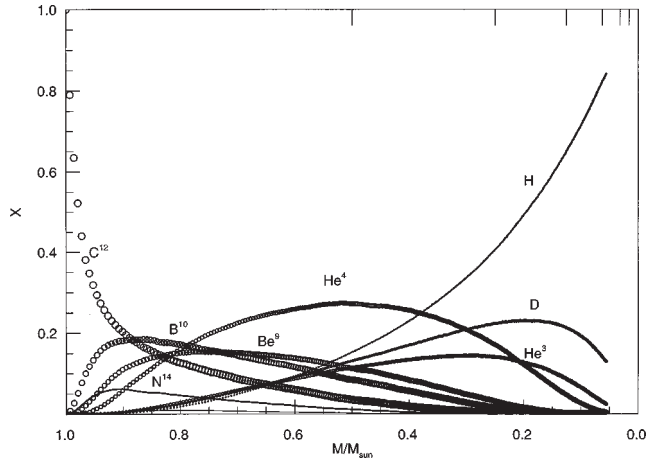

Figure 5.5 Chemical evolution of a white dwarf star during proton decay. The curves show the mass fractions of the major component nuclei in the star as a function of time, which is measured here in terms of the stellar mass. The initial state is a 1.0 $M_\odot$ white dwarf made of pure 12C. This simulation includes the effects of spallation and radioactivity (see text).

effects emerge: The first effect is that degeneracy is lifted and the star ceases to be a white dwarf. The second effect is that the object becomes optically thin to its internal radiation produced by proton decay and thus ceases to be a star. In the following discussion, we present simple estimates of the mass scales at which these events occur.

When the star has lost enough of its initial mass to become nondegenerate, most of the nucleons in the star will be in the form of hydrogen (see the previous section). A cold star composed of pure hydrogen will generally have a thick envelope of molecular hydrogen surrounding a degenerate core of atomic hydrogen. As the stellar mass continues to decline through the process of proton decay, the degenerate core becomes increasingly smaller and finally disappears altogether. This transition occurs when the degeneracy energy, the Coulomb energy, and the self-gravitational energy of the star are all comparable in magnitude; this event, in turn, occurs when the central pressure P_C drops below a critical value of roughly a few Megabars ($P_C \sim 10^{12}$ dyne/cm^2). The central pressure in a star can be written in the form

$$P_C = \beta \frac{GM_*^2}{R_*^4}, \tag{4.20}$$

where β is a dimensionless number of order unity. Using the white dwarf mass/radius relation in the form of equation [4.13] and setting $Z = A = 1$, we find the central pressure as a function of stellar mass,

$$P_C \approx \frac{\beta}{410} M_{\rm Pl}{}^{-10} m_e^4 m_P^{20/3} M_*^{10/3}, \tag{4.21}$$

or, equivalently (in cgs units),

$$P_C \approx 2\times 10^{21}\ \text{dyne/ cm}^2 \left(\frac{M_\star}{1M_\odot}\right)^{10/3}. \tag{4.22}$$

Combining these results, we find that the mass scale $M_{\star nd}$ at which the star becomes nondegenerate is given by

$$M_{\star \text{nd}} \approx 10^{-3} M_\odot. \tag{4.23}$$

This mass scale is roughly the mass of a giant planet such as Jupiter (for more detailed discussion of this issue, see also Hamada & Salpeter, 1961; Shu, 1982). At this point in its evolution, the star has a radius $R_\star \sim 0.1R_e \sim 7\times 10^9$ cm and a mean density of roughly $\rho \sim 1$ g/cm^3; these properties are also comparable to those of Jupiter. As a reference point, notice also that neutral hydrogen atoms packed into a cubic array with sides equal to one Bohr radius would give a density of 1.4 g/cm^3. At this transition, a star powered by proton decay has luminosity $L_\star \approx 10^{-27} L_e$ and effective surface temperature $T_\star \approx 0.0034$ K.

Once the star becomes nondegenerate, it follows new track in the H-R diagram. The expressions for the luminosity and surface temperature (see equations [4.8] and [4.12]) remain valid, but the mass/radius relation changes. Since the density of matter is determined by Coulomb forces for the small mass scales of interest, the density is roughly constant with a value $\rho_0 \sim 1$ g/cm^3. We can thus use the simple relationship $M_\star = 4\pi\rho_0 R_\star^3 / 3$. Combining these results, we obtain the relation

$$L_\star = \frac{36\pi}{F^2}\frac{\sigma_B^3}{\Gamma_P^2\rho_0^2} T_\star^{12}. \tag{4.24}$$

or, in terms of numerical values,

$$L_\star \approx 10^{-27} L_\odot \left[\frac{T_\star}{0.0034\,\text{K}}\right]^{12}. \tag{4.25}$$

This steep power-law implies that the effective temperature of the star does not change very much during the final phases of evolution (the mass has to decrease by 12 orders of magnitude in order for the temperature to change by a factor of 10). As a result, effective surface temperatures of order $T_\star \sim 10^{-3}$ K characterize the final phases of stellar evolution.

As the star loses mass, it also becomes increasingly optically thin to radiation. As an object becomes transparent, it becomes difficult to meaningfully consider the remnant as a star. An object becomes optically thin when

$$R,n\sigma < 1, \tag{4.26}$$

where n is the number density of targets and σ is the cross section of interaction between the radiation field and the stellar material. In this present context, we must consider whether the star is optically thin to both the gamma rays produced by proton decay and also to the internal radiation at longer wavelengths characteristic of its bolometric surface temperature. This latter condition is required for the radiation field to be thermalized.

We first consider the conditions for which the star becomes optically thin to the gamma rays (with energies $E\gamma \sim 250$ MeV) produced by proton decay. Since we are considering the interaction of gamma rays with matter, we can write the cross section in the form

$$\sigma = C\sigma_T = C\frac{8\pi}{3}\frac{\alpha^2}{m_e^2}, \tag{4.27}$$

where C is a dimensionless number (of order unity) and σ_T is the Thompson cross section. To a rough approximation, the density will be $\rho \sim 1$ g/cm^3 and hence the number density will have a roughly constant value $n \sim 10^{24}$ cm^{-3}. Using these values, we find that the "star" will be safely optically thick to gamma rays provided its characteristic size is larger than about one meter. In other words, the object must be as big as a large rock. These rocks will not, however, look very much like stars. At the extremely low bolometric temperatures characteristic of the stellar photospheres at these late times, the wavelength of the photospheric photons will be macroscopic and hence will interact much less strongly than the gamma rays. As a result, the spectral energy distribution of these objects will suffer severe departures from blackbody spectral shapes.

In order to consider the optical depth of the star to its internal radiation field, we rewrite the condition [4.26] using the relation $n\sigma = \rho\kappa$, where κ is the opacity. As derived above (equation [4.24]), the surface temperature is a slowly varying function in this final phase of evolution; as a result, the wavelength of photons in the stellar photosphere will be of order $\lambda \sim 100$ cm. The interaction of this radiation with the star depends on the chemical purity and the crystal-grain structure of the stellar material. We can obtain a very rough estimate of the opacity by scaling from known astrophysical quantities. For interstellar graphite, for example, the opacity at $\lambda = 100$ μm is roughly $\kappa \sim 1$ cm^2/g and scales with wavelength according to $\kappa \propto \lambda^{-2}$ (see Draine & Lee, 1984). We thus estimate that the opacity in the outer layers of the star/rock will be $\kappa \sim 10^{-8}$ cm^2/g. Thus, in order for the star to be optically thick to its internal radiation, its radius must be $R_* > 10^8$ cm, which corresponds to a mass scale of

$$M_{*\,\mathrm{thin}} \sim 10^{24}\,\mathrm{g}. \tag{4.28}$$

All of these values should be regarded as highly approximate.

From these results, the ultimate future of white dwarfs, and indeed our own sun, becomes clear: A white dwarf emerges from degeneracy as a pure sphere of hydrogen when the mass drops below $M_* \sim 10^{-3} M_e$. Finally, the remaining object becomes

transparent to its own internal radiation when its mass dwindles to $M_\star \sim 10^{24}$ g, and at this point it is no longer a star. Stellar evolution thus effectively comes to an end.

Just prior to the conclusion of stellar evolution, the white dwarf experiences about 2000 proton decay events per second and hence has a luminosity of $L_\star \sim 10^{-33} L_e \sim 4$ erg/s, and a temperature $T_\star \sim 10^{-3}$ K. The time at which this transition occurs is given by $\tau \sim 21\Gamma_P^{-1}$.

Given these results, we can now describe the complete evolution of a 1.0 M_e star (e.g., the Sun), from its birth to its death. The entire evolution of the such a star in the Hertzsprung-Russell diagram is plotted in Figure 5.6. The star first appears on the stellar birthline (Stahler, 1988) and then follows a pre-main sequence track onto the main sequence. After exhausting its available hydrogen, the star follows conventional post-main sequence evolution, including red giant, horizontal branch, red supergiant, and planetary nebula phases. The star then becomes a white dwarf with mass $M_\star \approx 0.5 M_e$ and cools along a constant radius track. The white dwarf spends many cosmological decades $\eta = 11 - 25$ near the center of the diagram ($L_\star = 10^{14}$ W; $T_\star = 63$ K), where the star is powered by annihilation of WIMPs accreted from the galactic halo. When the supply of WIMPs is exhausted, the star cools relatively quickly and obtains its luminosity from proton decay ($L_\star \approx 400$ W). The star then follows the evolutionary track in the lower right part of the diagram (with $L_\star \sim T_\star^{12/5}$) until mass-loss from proton decay causes the star to become nondegenerate. The star then becomes a rock-like object supported by Coulomb forces and follows a steeper track (with $L_\star \sim T_\star^{12}$) in the H-R diagram until it becomes optically thin. At this point, the object ceases to be a star and stellar evolution effectively comes to an end. During its entire lifetime, the Sun will span roughly 33 orders of magnitude in luminosity, 9 orders of magnitude in mass, and 8 orders of magnitude in surface temperature.

E Neutron stars powered by proton decay

The evolution of neutron stars powered by proton decay is qualitatively similar to that of white dwarfs. Since neutron stars are (roughly) the same mass as white dwarfs, and since proton decay occurs on the size scale of an individual nucleon, the luminosity of the neutron star is given by equations [4.8] and [4.9]. To leading order, the mass/radius relation for a neutron star is the same as that of white dwarfs with the electron mass m_e replaced by the neutron mass (see equations [4.13] and [4.14]). Neutron stars are thus ~2000 times smaller than white dwarfs of the same mass, and have appropriately warmer surface temperatures. Neutron stars undergoing nucleon decay follow a track in the H-R diagram given by

$$L_\star = 10^{-24} L_\odot \left[\frac{T_\star}{3\,\mathrm{K}} \right]^{12/5} . \tag{4.29}$$

The final phases of the life of a neutron star will differ from the case of a white dwarf. In particular, the neutrons in a neutron star come out of degeneracy in a somewhat different manner than the electrons in a white dwarf. Within a neutron star,

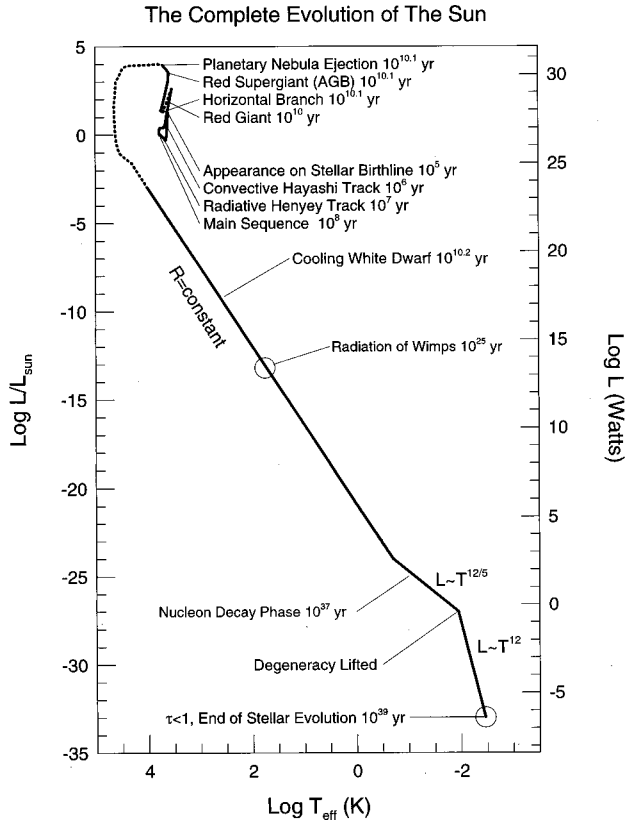

Figure 5.6 The complete evolution of the Sun (or any $1M_{\odot}$ star) in the H-R Diagram. The track shows the overall evolution of a star, from birth to final death. The star first appears in the H-R diagram on the stellar birthline and then follows a pre-main sequence track onto the main sequence. After its post-main sequence evolution (red giant, horizontal branch, red supergiant, and planetary nebula phases), the star becomes a white dwarf and cools along a constant radius track. The star spends many cosmological decades $\eta = 11-25$ at a point near the center of the diagram ($L_{*} = 1014$ W; $T_{*} = 63$ K), where the star is powered by annihilation of WIMPs accreted from the galactic halo. When the supply of WIMPs is exhausted, the star cools relatively quickly and obtains its luminosity from proton decay ($L_{*} \approx 400$ W). The star then follows the evolutionary track in the lower right part of the diagram (with $L_{*} \sim T_{*}^{12/5}$) until mass-loss from proton decay causes the star to become optically thin. At this point, the object ceases to be a star and stellar evolution comes to an end.

the neutrons exist and do not β-decay (into protons, electrons, and anti-neutrinos) because of the extremely high densities, which are close to nuclear densities in the stellar interior. On the exterior, however, every neutron star has a solid crust composed of ordinary matter. As a neutron star squanders its mass through nucleon decay, the radius swells and the density decreases. The outer layers of the star are less dense than

the central regions and hence the outer region will experience β-decay first. Thus, as the mass decreases, neutrons in the outer portion of the star begin to β-decay into their constituent particles and the star must readjust itself accordingly; the net effect is that the crust of ordinary matter thickens steadily and moves inwards towards the center. Once the stellar mass decreases below a critical value M_{C^*}, the crust reaches the center of the star and the transition becomes complete. At this point, the star will resemble a white dwarf more than a neutron star.

This process thus defines a minimum-mass neutron star (see Shapiro & Teukolsky, 1983), which is roughly characterized by the parameters

$$M_{C_*} = 0.0925 M_\odot, \quad \rho_C = 1.55 \times 10^{14}\,\mathrm{g\,cm^{-3}}, \quad R_* = 164\,\mathrm{km}, \tag{4.30}$$

where ρ_C is the central density of the star. It is hard to imagine current-day astrophysical processes which produce stellar objects near this limit. The transformation from a neutron star to a white dwarf occurs with a time scale given by

$$\tau = \frac{1}{\Gamma_P} \ln\left[M_0 / M_{C^*} \right] \approx \frac{2.7}{\Gamma_P}, \tag{4.31}$$

where $M_0 \approx 1.4 M_e$ is the initial mass of the neutron star. Notice that neutron stars have a possible mass range of only a factor of ~ 15, considerably smaller than the mass range available to white dwarfs.

F Higher-order proton decay

Not all particle physics theories predict proton decay through the process described above with decay rate Γ_P (equation [4.2] and Figure 5.4). In theories which do not allow proton decay through this first-order process, the proton can often decay through second-order processes and/or through gravitational effects. By a second-order process, we mean an interaction involving two protons and/or neutrons, i.e., $\Delta B = 2$, where B is the baryon number. The decay rate for these alternate decay channels is typically much smaller than that discussed above. In this section, we discuss the decay rates and time scales for these higher-order processes (see also Feinberg, Goldhaber, & Steigman, 1978; Wilczek & Zee, 1979; Weinberg, 1980; Mohapatra & Marshak, 1980).

We first consider a class of theories which allow baryon number violation, but do not have the proper vertices for direct proton decay ($\Delta B = 1$). In such theories, proton decay can sometimes take place through higher order processes ($\Delta B > 1$). For example, if the quarks in two nucleons interact as shown in Figure 5.7, the decay rate is roughly given by

$$\Gamma_2 \sim \alpha_5^4 \frac{m_P^9}{M_X^8}. \tag{4.32}$$

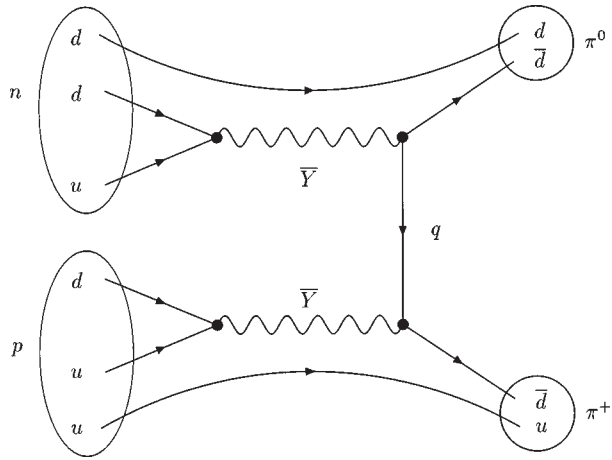

Figure 5.7 Representative Feynman diagram for nucleon decay for a B = 2 process, i.e., a decay involving two nucleons. The net result of this interaction (shown here in terms of the constituent quarks) is the decay of a neutron and a proton into two pions, n + p → π0 + π+. The Y particle mediates the baryon-number-violating process. Similar diagrams for neutron-neutron decay and for proton-proton decay can be obtained by changing the type of spectator quarks.

Even for this higher order example, the theory must have the proper vertices for this process to occur. We note that some theories forbid this class of decay channels and require $\Delta B = 3$ reactions in order for nucleon decay to take place (e.g., Goity & Sher, 1995; Castano & Martin, 1994). For the example shown in Figure 5.7, the decay rate is suppressed by a factor of $(m_p/M_X)^4 \sim 10^{64}$ relative to the simplest GUT decay channel. As a result, the time scale for proton decay through this second-order process is roughly given by

$$\tau_{P2} \approx 10^{101}\,\mathrm{yr}\left[\frac{M_X}{10^{16}\,\mathrm{GeV}}\right]^8, \tag{4.33}$$

and the corresponding cosmological time scale is

$$\eta_{P2} = 101 + 8\log_{10}\left[M_X / 10^{16}\,\mathrm{GeV}\right]. \tag{4.34}$$

In order for this decay process to take place, the protons involved must be near each other. For the case of interest, the protons in white dwarfs are (mostly) in carbon nuclei and hence meet this requirement. Similarly, the neutrons in a neutron star are all essentially at nuclear densities. Notice, however, that free protons in interstellar or intergalactic space will generally not decay through this channel.

The proton can also decay through virtual black-hole processes in quantum gravity theories (e.g., Zel'dovich, 1976; Hawking, Page, & Pope, 1979; Page, 1980; Hawking,

1987). Unfortunately, the time scale associated with this process is not very well determined, but it is estimated to lie in the range

$$10^{46}\ \text{yr} < \tau_{PBH} < 10^{169}\ \text{yr}\ , \tag{4.35}$$

with the corresponding range of cosmological decades

$$46 < \eta_{PBH} < 169\ . \tag{4.36}$$

Thus, within the (very large) uncertainty, this time scale for proton decay is commensurate with the second-order GUT processes discussed above.

We note that many other possible modes of nucleon decay exist. For example, supersymmetric theories can give rise to a double neutron decay process of the form shown in Figure 5.8a (see Goity & Sher, 1995). In this case, two neutrons decay into two neutral kaons. Within the context of standard GUTs, decay channels involving higher order diagrams can also occur. As another example, the process shown in Figure 5.8b involves three intermediate vector bosons and thus leads to a proton lifetime approximately given by

$$\eta_{P3} = 165 + 12\log_{10}\left[M_X / 10^{16}\,\text{GeV}\right]. \tag{4.37}$$

Other final states are possible (e.g., three pions), although the time scales should be comparable. This process (Figure 5.8b) involves only the most elementary baryon number violating processes, which allow interactions of the general form $qq \rightarrow q\bar{q}$. As a result, this decay mode is likely to occur even when the lower-order channels are not allowed.

Finally, we mention the case of sphalerons, which provide yet another mechanism that can lead to baryon number violation and hence proton decay. The vacuum structure of the electroweak theory allows for the non-conservation of baryon number; tunneling events between the different vacuum states in the theory give rise to a change in baryon number (for further details, see Rajaraman, 1987; Kolb & Turner, 1990). Because these events require quantum tunneling, the rate for this process is exponentially suppressed at zero temperature by the large factor $f = \exp[4\pi/\alpha_W] \sim 10^{172}$, where αW is the fine-structure constant for weak interactions. In terms of cosmological decades, the time scale for proton decay through this process has the form $\eta_P = \eta_0 + 172$, where η_0 is the natural time scale (for no suppression). Using the light crossing time of the proton to determine the natural time scale (i.e., we optimistically take $\eta_0 = -31$), we obtain the crude estimate $\eta_P \approx 141$. Since this time scale is much longer than the current age of the universe, this mode of proton decay has not been fully explored. In addition, this process has associated selection rules (e.g., 't Hooft, 1976) that place further limits on the possible events which exhibit nonconservation of baryon number. However, this mode of baryon number violation could play a role in the far future of the universe.

To summarize this discussion, we stress that many different mechanisms for baryon number violation and proton decay can be realized within modern theories of particle

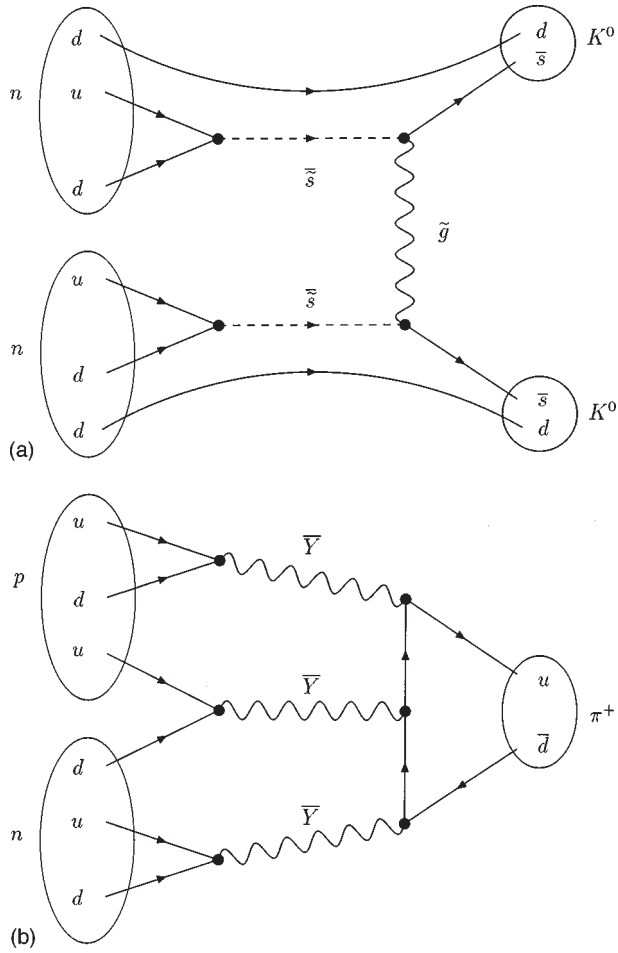

Figure 5.8 Representative Feynman diagram for higher order nucleon decay processes, shown here in terms of the constituent quarks. (a) Double neutron decay for a supersymmetric theory. The net reaction converts two neutrons *n* into two neutral kaons K^0. The tildes denote the supersymmetric partners of the particles. (b) Double nucleon decay involving three intermediate vector bosons Y. Other final states are possible (e.g., three pions), but the overall decay rate is comparable and implies a decay time scale $\eta P \sim 165 + 12 \log10[MY /1016GeV]$.

physics. As a result, it seems likely that the proton must eventually decay with a lifetime somewhere in the range

$$32 < \eta_P < 200\,, \tag{4.38}$$

where the upper bound was obtained by using $M_X \sim M_{PI}$ in equation [4.37].

To put these very long time scales in perspective, we note that the total number N_N of nucleons in the observable universe (at the present epoch) is roughly $N_N \sim 10^{78}$. Thus, for a decay time of $\eta = 100$, the expected number N_D of nucleons that have decayed within our observable universe during its entire history is far less than unity, $N_D \sim 10^{-12}$. The experimental difficulties involved in detecting higher order proton decay processes thus become clear.

If the proton decays with a lifetime corresponding to $\eta \sim 100 - 200$, the evolution of white dwarfs will be qualitatively the same as the scenario outlined above, but with a few differences. Since the evolutionary time scale is much longer, pycnonuclear reactions will be much more effective at building the chemical composition of the stars back up to nuclei of high atomic number. Thus, stars with a given mass will have higher atomic numbers for their constituent nuclei. However, the nuclear reaction rate (equation [4.17]) has an exponential sensitivity to the density. As the star loses mass and becomes less dense (according to the white dwarf mass/radius relation [4.13, 4.14]), pycnonuclear reactions will shut down rather abruptly. If these nuclear reactions stop entirely, the star would quickly become pure hydrogen and proton decay through a two-body process would be highly suppressed. However, hydrogen tends to form molecules at these extremely low temperatures. The pycnonuclear reaction between the two protons in a hydrogen molecule proceeds at a fixed rate which is independent of the ambient conditions and has a time scale of roughly $\eta \approx 60$ (see Dyson, 1979, Shapiro & Teukolsky, 1983, and §III.C for simple estimates of pycnonuclear reaction rates). This reaction will thus convert the star into deuterium and helium on a time scale significantly shorter than that of higher order proton decay. The resulting larger nuclei can then still decay through a second- or third-order process. We also note that this same mechanism allows for hydrogen molecules in intergalactic space to undergo proton decay through a two-body process.

G Hawking radiation and the decay of black holes

Black holes cannot live forever; they evaporate on long time scales through a quantum mechanical tunneling process that produces photons and other products (Hawking, 1975). In particular, black holes radiate a thermal spectrum of particles with an effective temperature given by

$$T_{BH} = \frac{1}{8\pi G M_{BH}}, \tag{4.39}$$

where M_{BH} is the mass of the black hole. The total life time of the black hole thus becomes

$$\tau_{BH} = \frac{2560\pi}{g_*} G^2 M_{BH}^3, \tag{4.40}$$

where g_* determines the total number of effective degrees of freedom in the radiation field. Inserting numerical values and scaling to a reference black hole mass of $10^6\,M_e$, we find the time scale

$$\tau_{BH} = 10^{83}\ \mathrm{yr}\left[M_{BH}/10^6 M_\odot\right]^3 . \tag{4.41}$$

or, equivalently,

$$\eta_{BH} = 83 + 3\log_3\left[M_{BH}/10^6 M_\odot\right] \tag{4.42}$$

Thus, even a black hole with a mass comparable to a galaxy ($M_{BH} \sim 10^{11} M_e$) will evaporate through this process on the time scale $\eta_{BH} \sim 98$. One important consequence of this result is that for $\eta > 100$, a large fraction of the universe will be in the form of radiation, electrons, positrons, and other decay products.

H Proton decay in planets

Planets will also eventually disintegrate through the process of proton decay. Since nuclear reactions have a time scale ($\eta \sim 1500$) much longer than that of proton decay and hence are unimportant (see Dyson, 1979), the chemical evolution of the planet is well described by a simple proton decay cascade scenario (see §IV.C). In particular, this cascade will convert a planet initially composed of iron into a hydrogen lattice in ~ 6 proton half-lives, or equivalently, on a time scale given by

$$\tau_{\mathrm{planet}} \approx \frac{6\ln 2}{\Gamma_P} \approx 10^{38}\ \mathrm{yr}; \quad \eta_{\mathrm{planet}} \approx 38\,, \tag{4.43}$$

This time scale also represents the time at which the planet is effectively destroyed.

During the epoch of proton decay, planets radiate energy with an effective luminosity given by

$$L_{\mathrm{planet}} = F M_{\mathrm{planet}}(t)\Gamma_P \approx 10^{-30} L_\odot \left[\frac{M_{\mathrm{planet}}}{M_E}\right], \tag{4.44}$$

where M_E is the mass of the Earth and where we have used a proton decay lifetime of 10^{37} yr. The efficiency factor F is expected to be of order unity. Thus, the luminosity corresponds to ~ 0.4 mW.

V Long-term evolution of the universe

In spite of the wealth of recent progress in our understanding of cosmology, the future evolution of the universe cannot be unambiguously predicted. In particular,

the geometry of the universe as a whole remains unspecified. The universe can be closed ($k = +1$; $\Omega > 1$), flat ($k = 0$; $\Omega = 1$), or open ($k = -1$; $\Omega < 1$). In addition, the contribution of vacuum energy density remains uncertain and can have important implications for the long-term evolution of the universe.

A Future expansion of a closed universe

If the universe is closed, then the total lifetime of the universe, from Big Bang to Big Crunch, can be relatively short in comparison with the characteristic time scales of many of the physical processes considered in this paper. For a closed universe with density parameter $\Omega_0 > 1$, the total lifetime τ_U of the universe can be written in the form

$$\tau_U = \Omega_0 \left(\Omega_0 - 1\right)^{3/2} \pi H_0^{-1} , \tag{5.1}$$

where H_0 is the present value of the Hubble parameter (see, e.g., Peebles, 1993). Notice that, by definition, the age $\tau_U \to \infty$ as $\Omega_0 \to 1$. Current cosmological observations suggest that the Hubble constant is restricted to lie in the range 50 – 100 km s^{-1} Mpc^{-1} (e.g., Riess, Press, & Kirshner, 1995), and hence the time scale H_0^{-1} is restricted to be greater than ~ 10 Gyr. Additional observations (e.g., Loh & Spillar, 1986) suggest that $\Omega_0 < 2$. Using these results, we thus obtain a lower bound on the total lifetime of the universe,

$$\tau_U > 20\pi \text{ Gyr} . \tag{5.2}$$

In terms of the time variable η, this limit takes the form

$$\eta_U > 10.8 . \tag{5.3}$$

This limit is not very strong – if the universe is indeed closed, then there will be insufficient time to allow for many of the processes we describe in this paper.

We also note that a closed-universe model can in principle be generalized to give rise to an oscillating universe. In this case, the Big Crunch occurring at the end of the universe is really a "Big Bounce" and produces a new universe of the next generation. This idea originated with Lemaître (1933) and has been subsequently considered in many different contexts (from Tolman, 1934 to Peebles, 1993).

B Density fluctuations and the expansion of a flat or open universe

The universe will either continue expanding forever or will collapse back in on itself, but it is not commonly acknowledged that observations are unable to provide a definitive answer to this important question. The goal of many present-day astronomical observations is to measure the density parameter Ω, which is the ratio of the density

of the universe to that required to close the universe. However, measurements of Ω do not necessarily determine the long-term fate of the universe.

Suppose, for example, that we can ultimately measure Ω to be some value Ω_0 (either less than or greater than unity). This value of Ω_0 means that the density within the current horizon volume has a given ratio to the critical density. If we could view the universe (today) on a much larger size scale (we can't because of causality), then the mean density of the universe of that larger size scale need not be the same as that which we measure within our horizon today. Let Ω_{big} denote the ratio of the density of the universe to the critical density on the aforementioned larger size scale. In particular, we could measure a value $\Omega_0 < 1$ and have $\Omega_{big} > 1$, or, alternately, we could measure $\Omega_0 > 1$ and have $\Omega_{big} < 1$. This possibility has been discussed at some length by Linde (1988, 1989, 1990).

To fix ideas, consider the case in which the local value of the density parameter is $\Omega_0 \approx 1$ and the larger scale value is $\Omega_{big} = 2 > 1$. (Note that Ω is not constant in time and hence this value refers to the time when the larger scale enters the horizon.) In other words, we live in an apparently flat universe, which is actually closed on a larger scale. This state of affairs requires that our currently observable universe lies within a large-scale density fluctuation of amplitude

$$\frac{\Delta\rho}{\rho} = \frac{\Omega_0 - \Omega_{\rm big}}{\Omega_{\rm big}} = -\frac{1}{2}, \tag{5.4}$$

where the minus sign indicates that we live in a locally underdense region. Thus, a density perturbation with amplitude of order unity is required; furthermore, as we discuss below, the size scale of the perturbation must greatly exceed the current horizon size.

On size scales comparable to that of our current horizon, density fluctuations are constrained to be quite small ($\Delta\rho/\rho \sim 10^{-5}$) because of measurements of temperature fluctuations in the cosmic microwave background radiation (Smoot et al., 1992; Wright et al., 1992). On smaller size scales, additional measurements indicate that density fluctuations are similarly small in amplitude (e.g., Meyer, Cheng, & Page, 1991; Gaier et al., 1992; Schuster et al., 1993). The microwave background also constrains density fluctuations on scales *larger than the horizon* (e.g., Grischuk & Zel'dovich, 1978), although the sensitivity of the constraint decreases with increasing size scale λ according to the relation $\sim (\lambda_{hor}/\lambda)^2$, where λ_{hor} is the horizon size. Given that density fluctuations have amplitudes of roughly $\sim 10^{-5}$ on the size scale of the horizon today, the smallest size scale λ_1 for which fluctuations can be of order unity is estimated to be

$$\lambda_1 \sim 300\lambda_{\rm hor} \approx 10^6\ {\rm Mpc}\,. \tag{5.5}$$

For a locally flat universe ($\Omega_0 \approx 1$), density fluctuations with this size scale will enter the horizon at a time $t_1 \approx 3 \times 10^7\ t_0 \approx 3 \times 10^{17}$ yr, or, equivalently, at the cosmological decade

$$\eta_1 \approx 17.5 \,. \tag{5.6}$$

This time scale represents a lower bound on the (final) age of the universe if the present geometry is spatially flat. In practice, the newly closed universe will require some additional time to re-collapse (see equation [5.1]) and hence the lower bound on the total age becomes approximately $\eta > 18$.

The situation is somewhat different for the case of an open universe with $\Omega_0 < 1$. If the universe is open, then the expansion velocity will (relatively) quickly approach the speed of light, i.e., the scale factor will expand according to $R \propto t$ (for this discussion, we do not include the possibility that $\Omega_0 = 1 - \varepsilon$, where $\varepsilon \ll 1$, i.e., we consider only manifestly open cases). In this limit, the (comoving) particle horizon expands logarithmically with time and hence continues to grow. However, the speed-of-light sphere – the distance out to which particles in the universe are receding at the speed of light – approaches a constant in comoving coordinates. As a result, density perturbations on very large scales will remain effectively "frozen out" and are thus prevented from further growth as long as the universe remains open. Because the comoving horizon continues to grow, albeit quite slowly, the possibility remains for the universe to become closed at some future time. The logarithmic growth of the horizon implies that the time scale for the universe to become closed depends exponentially on the size scale λ_1 for which density perturbations are of order unity. The resulting time scale is quite long ($\eta \gg 100$), even compared to the time scales considered in this paper.

To summarize, if the universe currently has a nearly flat spatial geometry, then microwave background constraints imply a lower bound on the total age of universe, $\eta > 18$. The evolution of the universe at later times depends on the spectrum of density perturbations. If large-amplitude perturbations ($\Delta\rho/\rho > 1$) enter the horizon at late times, then the universe could end in a big crunch at some time $\eta > \eta_1 = 17.5$. On the other hand, if the very large-scale density perturbations have small amplitude ($\Delta\rho / \rho \ll 1$), then the universe can continue to expand for much longer time scales. If the universe is currently open, then large-scale density perturbations are essentially frozen out.

C Inflation and the future of the universe

The inflationary universe scenario was originally invented (Guth, 1981) to solve the horizon problem and the flatness problem faced by standard Big Bang cosmology (see also Albrecht & Steinhardt, 1982; Linde, 1982). The problem of magnetic monopoles was also a motivation, but will not be discussed here. In addition, inflationary models which utilize "slowly rolling" scalar fields can produce density fluctuations which later grow into the galaxies, clusters, and super-clusters that we see today (e.g., Bardeen, Steinhardt, & Turner, 1983; Starobinsky, 1982; Guth & Pi, 1982; Hawking, 1982).

During the inflationary epoch, the scale factor of the universe grows superluminally (usually exponentially with time). During this period of rapid expansion, a small causally connected region of the universe inflates to become large enough to contain

the presently observable universe. As a result, the observed homogeneity and isotropy of the universe can be explained, as well as the observed flatness. In order to achieve this resolution of the horizon and flatness problems, the scale factor of the universe must inflate by a factor of e^{N_I}, where the number of e-foldings $N_I \sim 60$. At the end of this period of rapid expansion, the universe must be re-thermalized in order to become radiation dominated and recover the successes of standard Big Bang theory.

Since the conception of inflation, many models have been produced and many treatments of the requirements for sufficient inflation have been given (e.g., Steinhardt & Turner, 1984; Kolb & Turner, 1990; Linde, 1990). These constraints are generally written in terms of explaining the flatness and causality of the universe at the present epoch. However, it is possible, or even quite likely, that inflation will solve the horizon and flatness problems far into the future. In this discussion, we find the number N_I of inflationary e-foldings required to solve the horizon and flatness problems until a future cosmological decade η.

Since the number of e-foldings required to solve the flatness problem is (usually) almost the same as that required to solve the horizon problem, it is sufficient to consider only the latter (for further discussion of this issue, see, e.g., Kolb & Turner, 1990; Linde, 1990). The condition for sufficient inflation can be written in the form

$$\frac{1}{(HR)_\eta} < \frac{1}{(HR)_B}, \tag{5.7}$$

where the left hand side of the inequality refers to the inverse of the product of the Hubble parameter and the scale factor evaluated at the future cosmological decade η and the right hand side refers to the same quantity evaluated at the beginning of the inflationary epoch.

The Hubble parameter at the beginning of inflation takes the form

$$H_B^2 = \frac{8\pi}{3}\frac{M_I^4}{M_{\mathrm{Pl}}^2}, \tag{5.8}$$

where M_I is the energy scale at the start of inflation (typically, the energy scale $M_I \sim 10^{16}$ GeV, which corresponds to cosmological decade $\eta_I \sim -44.5$). Similarly, the Hubble parameter at some future time η can be written in the form

$$H_\eta^2 = \frac{8\pi}{3}\frac{M_\eta^4}{M_{\mathrm{Pl}}^2}, \tag{5.9}$$

where the energy scale $M\eta$ is defined by

$$\rho(\eta) \equiv M_\eta^4 = \rho_0 R_\eta^{-3}. \tag{5.10}$$

In the second equality, we have written the energy density in terms of its value ρ_0 at the present epoch and we assume that the universe remains matter dominated. We

also assume that the evolution of the universe is essentially adiabatic from the end of inflation (scale factor R_{end}) until the future epoch of interest (scale factor $R\eta$), i.e.,

$$\frac{R_{end}}{R_\eta} = \frac{T_\eta}{fM_I}, \tag{5.11}$$

where $T\eta = T_0/R\eta$ is the CMB temperature at time η and $T_0 \approx 2.7$ K is the CMB temperature today. The quantity fM_I is the CMB temperature at the end of inflation, after thermalization, and we have introduced the dimensionless factor $f < 1$.

Combining all of the above results, we obtain the following constraint for sufficient inflation,

$$e^{N_I} = \frac{R_{end}}{R_B} > \frac{M_I T_0 R_\eta^{1/2}}{f\sqrt{\rho_0}}. \tag{5.12}$$

Next, we write the present day energy density ρ_0 in terms of the present day CMB temperature T_0,

$$\rho_0 = \beta^2 T_0^4, \tag{5.13}$$

where $\beta \approx 100$. The number of e-foldings is thus given by

$$N_I = \ln\left[R_{end}/R_B\right] = \ln\left[M_I/\beta T_0\right] + \frac{1}{2}\ln R_\eta - \ln f. \tag{5.14a}$$

Inserting numerical values and using the definition [1.1] of cosmological decades, we can write this constraint in the form

$$N_I \approx 61 + \ln\left[M_I/\left(10^{16}\,\mathrm{GeV}\right)\right] + \frac{1}{3}(\eta - 10)\ln 10. \tag{5.14b}$$

For example, in order to have enough inflation for the universe to be smooth and flat up to the cosmological decade $\eta = 100$, we require $N_I \approx 130$ e-foldings of inflation. This value is not unreasonable in that $N_I = 130$ is just as natural from the point of view of particle physics as the $N_I = 61$ value required by standard inflation.

We must also consider the density perturbations produced by inflation. All known models of inflation produce density fluctuations and most models predict that the amplitudes are given by

$$\frac{\Delta\rho}{\rho} \approx \frac{1}{10}\frac{H^2}{\dot{\Phi}}, \tag{5.15}$$

where H is the Hubble parameter and Φ is the scalar field responsible for inflation (Starobinsky, 1982; Guth & Pi, 1982; Hawking, 1982; Bardeen, Steinhardt, & Turner, 1983). In models of inflation with more than one scalar field (e.g., La & Steinhardt,

1989; Adams & Freese, 1991), the additional fields can also produce density fluctuations in accordance with equation [5.15].

In order for these density fluctuations to be sufficiently small, as required by measurements of the cosmic microwave background, the potential $V(\Phi)$ for the inflation field must be very flat. This statement can be quantified by defining a "fine-tuning parameter" λ_{FT} through the relation

$$\lambda_{FT} \equiv \frac{\Delta V}{\left(\Delta\Phi\right)^4}\,, \tag{5.16}$$

where ΔV is the change in the potential during a given portion of the inflationary epoch and $\Delta\Phi$ is the change in the scalar field over the same period (Adams, Freese, & Guth, 1991). The parameter λ_{FT} is constrained to less than $\sim 10^{-8}$ for all models of inflation of this class and is typically much smaller, $\lambda_{FT} \sim 10^{-12}$, for specific models. The required smallness of this parameter places tight constraints on models of inflation.

The aforementioned constraints were derived by demanding that the density fluctuations (equation [5.15]) are sufficiently small in amplitude over the size scales of current cosmological interest, i.e., from the horizon size (today) down to the size scale of galaxies. These density perturbations are generated over $N\delta \approx 8$ e-foldings during the inflationary epoch. However, as discussed in §V.B, large-amplitude density fluctuations can come across the horizon in the future and effectively close the universe (see also Linde, 1988, 1989, 1990). In order for the universe to survive (not become closed) up until some future cosmological decade η, density fluctuations must be small in amplitude for all size scales up to the horizon size at time η (within an order of magnitude – see equation [5.1]). As a result, inflation must produce small amplitude density fluctuations over many more e-foldings of the inflationary epoch, namely

$$N_\delta \approx 8 + \frac{1}{3}\left(\eta - 10\right)\ln 10\,, \tag{5.17}$$

where η is the future cosmological decade of interest. For example, for $\eta = 100$ we would require $N\delta \approx 77$. Although this larger value of $N\delta$ places a tighter bound on the fine-tuning parameter λ_{FT}, and hence a tighter constraint on the inflationary potential, such bounds can be accommodated by inflationary models (see Adams, Freese, & Guth, 1991 for further discussion). Loosely speaking, once the potential is flat over the usual $N\delta = 8$ e-foldings required for standard inflationary models, it is not that difficult to make it flat for $N\delta = 80$.

D Background radiation fields

Many of the processes discussed in this paper will produce background radiation fields, which can be important components of the universe (see, e.g., Bond, Carr, & Hogan, 1991 for a discussion of present day backgrounds). Stars produce radiation fields and low-mass stars will continue to shine for several more cosmological decades

(§II). The net effect of WIMP capture and annihilation in white dwarfs (§III.E) will be to convert a substantial portion of the mass energy of galactic halos into radiation. Similarly, the net effect of proton decay (§IV) will convert the mass energy of the baryons in the universe into radiation. Finally, black holes will evaporate as well, (§IV.H), ultimately converting their rest mass into radiation fields. As we show below, each of these radiation fields will dominate the radiation background of the universe for a range of cosmological decades, before being successively redshifted to insignificance.

The overall evolution of a radiation field in an expanding universe can be described by the simple differential equation,

$$\frac{d\rho_{\rm rad}}{dt} + 4\frac{\dot{R}}{R}\rho_{\rm rad} = S(t)\,, \tag{5.18}$$

where ρ_{rad} is the energy density of the radiation field and $S(t)$ is a source term (see, e.g., Kolb & Turner, 1990).

Low-mass stars will continue to shine far into the future. The source term for this stellar radiation can be written in the form

$$S_\star(t) = n_\star L_\star = \varepsilon_\star \Omega_\star \rho_0 R^{-3}\frac{1}{t_\star}\,, \tag{5.19}$$

where $L_\star$ and $n_\star$ are the luminosity and number density of the low-mass stars. In the second equality, we have introduced the present day mass fraction of low-mass stars $\Omega_\star$, the nuclear burning efficiency $\varepsilon_\star \sim 0.007$, the effective stellar lifetime $t_\star$, and the present day energy density of the universe ρ_0. For this example, we have written these expressions for a population of stars with only a single mass; in general, one should of course consider a distribution of stellar masses and then integrate over the distribution. As a further refinement, one could also include the time dependence of the stellar luminosity $L_\star$ (see §II).

For a given geometry of the universe, we find the solution for the background radiation field from low-mass stars,

$$\rho_{\rm rad,\star} = \varepsilon_\star \Omega_\star \rho(R) f\frac{t}{t_\star}\,, \tag{5.20}$$

where the dimensionless factor $f = 1/2$ for an open universe and $f = 3/5$ for a flat universe. This form is valid until the stars burn out at time $t = t_\star$. After that time, the radiation field simply redshifts in the usual manner, $\rho_{rad\star} \sim R^{-4}$.

For the case of WIMP annihilation in white dwarfs, the source term is given by

$$S_W(t) = L_\star n_\star = \Omega_W \rho_0 R^{-3}\Gamma\,, \tag{5.21}$$

where $L_\star$ and $n_\star$ are the luminosity and number density of the white dwarfs. In the second equality, we have written the source in terms of the energy density in WIMPs,

where Ω_W is the present day mass fraction of WIMPs and Γ is the effective annihilation rate. The solution for the background radiation field from WIMP annihilation can be found,

$$\rho_{\rm wrb}(t) = f\Omega_W \rho(R)\Gamma t \ , \tag{5.22}$$

where the dimensionless factor f is defined above. This form is valid until the galactic halos begin to run out of WIMP dark matter at time $t \sim \Gamma^{-1} \sim 10^{25}$ yr, or until the galactic halo ejects nearly all of its white dwarfs. We note that direct annihilation of dark matter will also contribute to the background radiation field of the universe. However, this radiation will be highly nonthermal; the annihilation products will include gamma rays with characteristic energy $E\gamma \sim 1$ GeV.

For the case of proton decay, the effective source term for the resulting radiation field can be written

$$S_P(t) = F\Omega_B \rho_0 R^{-3} \Gamma_P \, e^{-\Gamma_P t} \ , \tag{5.23}$$

where Ω_B is the present day contribution of baryons to the total energy density ρ_0, Γ_P is the proton decay rate, and F is an efficiency factor of order unity. For a given geometry of the universe, we obtain the solution for the background radiation field from proton decay,

$$\rho_{\rm prb}(t) = F\Omega_B \rho(R) F(\xi) \ , \tag{5.24}$$

where $F(\xi)$ is a dimensionless function of the dimensionless time variable $\xi \equiv \Gamma_P t$. For an open universe,

$$F(\xi) = \frac{1-(1+\xi)e^{-\xi}}{\xi} \ , \tag{5.25}$$

whereas for a flat universe,

$$F(\xi) = -\xi^{-2/3} \int_0^{\xi} x^{2/3} e^{-x} \, dx = \xi^{-2/3} \gamma(5/3, \xi) \ , \tag{5.26}$$

where $\gamma(5/3, \xi)$ is the incomplete gamma function (Abramowitz & Stegun, 1972).

For black hole evaporation, the calculation of the radiation field is more complicated because the result depends on the mass distribution of black holes in the universe. For simplicity, we will consider a population of black holes with a single mass M and mass fraction Ω_{BH} (scaled to the present epoch). The source term for black hole evaporation can be written in the form

$$S_{BH}(t) = \Omega_{BH} \rho_0 R^{-3} \frac{1}{3\tau_{BH}} \frac{1}{1 - t/\tau_{BH}} \ , \tag{5.27}$$

where τ_{BH} is the total lifetime of a black hole of the given mass M (see equation [4.37]). For an open universe, we obtain the solution for the background radiation field from black hole evaporation

$$\rho_{\text{bhr}}(t) = \Omega_{BH}\rho(R)F(\xi), \tag{5.28}$$

where the dimensionless time variable $\xi = t/\tau_{BH}$. For an open universe, the dimensionless function $F(\xi)$ is given by

$$F(\xi) = \frac{1}{3\xi}\left\{\ln\left[\frac{1}{1-\xi}\right] - \xi\right\}, \tag{5.29}$$

whereas for a flat universe,

$$F(\xi) = \frac{1}{3\xi^{2/3}}\int_0^{\xi}\frac{x^{2/3}dx}{1-x} \tag{5.30}$$

Each of the four radiation fields discussed here has the same general time dependence. For times short compared to the depletion times, the radiation fields have the form

$$\rho(t) \approx \Omega_X\rho(R)\Gamma_X t, \tag{5.31}$$

where Ω_X is the present day abundance of the raw material and Γ_X is the effective decay rate (notice that we have neglected dimensionless factors of order unity). After the sources (stars, WIMPs, protons, black holes) have been successively exhausted, the remaining radiation fields simply redshift away, i.e.,

$$\rho(t) = \rho(t_{\text{end}})(R/R_{\text{end}})^{-4}, \tag{5.32}$$

where the subscript refers to the end of the time period during which the ambient radiation was produced.

Due to the gross mismatch in the characteristic time scales, each of the radiation fields will provide the dominate contribution to the radiation content of the universe over a given time period. This trend is illustrated in Figure 5.9, which shows the relative contribution of each radiation field as a function of cosmological time η. For purposes of illustration, we have assumed an open universe and the following source abundances: low-mass stars $\Omega_* = 10^{-3}$, weakly interacting massive particles $\Omega_W = 0.2$, baryons $\Omega_B = 0.05$, and black holes $\Omega_{BH} = 0.1$. At present, the cosmic microwave background (left over from the big bang itself) provides the dominant radiation component. The radiation field from star light will dominate the background for the next several cosmological decades. At cosmological decade $\eta \sim 16$, the radiation field resulting from WIMP annihilation will overtake the starlight background and become the dominant component. At the cosmological decade $\eta \sim 30$, the WIMP annihilation radiation field will have redshifted away and the radiation field from proton decay will begin to dominate. At much longer time scales, $\eta \sim 60$, the radiation field from black hole evaporation provides the dominant contribution (where we have used 10^6 M_e black holes for this example).

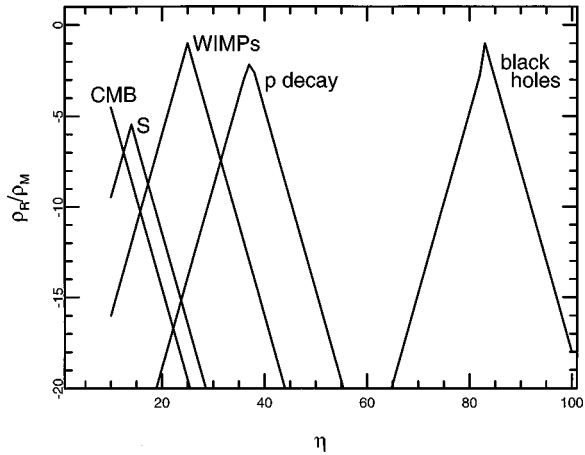

Figure 5.9 Background radiation fields in the universe. The vertical axis represents the ratio of the energy density in radiation to the total energy density (assuming the universe remains matter dominated). The horizontal axis is given in terms of cosmological decades η. The various curves represent the radiation fields from the cosmic microwave background (CMB), light from low-mass stars (S), radiation from WIMP annihilation in white dwarfs (WIMPs), radiation from proton decay (p decay), and black hole evaporation (black holes).

The discussion thus far has focused on the total energy density ρ_{rad} of the background radiation fields. One can also determine the spectrum of the background fields as a function of cosmological time, i.e., one could follow the time evolution of the radiation energy density per unit frequency. In general, the spectra of the background radiation fields will be non-thermal for two reasons:

1. The source terms are not necessarily perfect blackbodies. The stars and black holes themselves produce nearly thermal spectra, but objects of different masses will radiate like blackbodies of different temperatures. One must therefore integrate over the mass distribution of the source population. It is interesting that this statement applies to all of the above sources. For the first three sources (low-mass stars, white dwarfs radiating WIMP annihilation products, and white dwarfs powered by proton decay), the mass distribution is not very wide and the resulting composite spectrum is close to that of a blackbody. For the case of black holes, the spectrum is potentially much wider, but the mass distribution is far more uncertain.
2. The expansion of the universe redshifts the radiation field as it is produced and thereby makes the resultant spectrum wider than a thermal distribution. However, due to the linear time dependence of the emission (equation [5.31]), most of the radiation is emitted in the final cosmological decade of the source's life. The redshift effect is thus not as large as one might naively think.

To summarize, the radiation fields will experience departures from a purely thermal distribution. However, we expect that the departures are not overly severe.

The above results, taken in conjunction with our current cosmological understanding, imply that it is unlikely that the universe will become radiation dominated in the far future. The majority of the energy density at the present epoch is (most likely) in the form of non-baryonic dark matter of some kind. A substantial fraction of this dark matter resides in galactic halos, and some fraction of these halos can be annihilated and hence converted into radiation through the white dwarf capture process outlined in §III.E. However, an equal or larger fraction of this dark matter resides outside of galaxies and/or can escape destruction through evaporation from galactic halos. Thus, unless the dark matter particles themselves decay into radiation, it seems that enough non-baryonic dark matter should survive to keep the universe matter dominated at all future epochs; in addition, the leftover electrons and positrons will help prevent the universe from becoming radiation-dominated (see also Page & McKee, 1981ab).

E Possible effects of vacuum energy density

If the universe contains a nonvanishing contribution of vacuum energy to the total energy density, then two interesting long-term effects can arise. The universe can enter a second inflationary phase, in which the universe expands superluminally (Guth, 1981; see also Albrecht & Steinhardt, 1983; Linde, 1982). Alternately, the vacuum can, in principle, be unstable and the universe can tunnel into an entirely new state (e.g., Coleman, 1977, 1985). Unfortunately, the contribution of the vacuum to the energy density of the universe remains unknown. In fact, the "natural value" of the vacuum energy density appears to be larger than the cosmologically allowed value by many orders of magnitude. This discrepancy is generally known as the "cosmological-constant problem" and has no currently accepted resolution (see the reviews of Weinberg, 1989; Carroll, Press, & Turner, 1992).

1 Future inflationary epochs

We first consider the possibility of a future inflationary epoch. The evolution equation for the universe can be written in the form

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{8\pi G}{3}\left(\rho_M + \rho_{\rm vac}\right), \tag{5.33}$$

where R is the scale factor, ρ_M is the energy density in matter, and ρ_{vac} is the vacuum energy density. We have assumed a spatially flat universe for simplicity. The matter density varies with the scale factor according to $\rho_M \sim R^{-3}$, whereas the vacuum energy density is constant. We can define the ratio

$$\nu \equiv \rho_{\rm vac} / \rho_0 , \tag{5.34}$$

i.e., the ratio of the vacuum energy density to that of the matter density ρ_0 at the present epoch. We can then integrate equation [5.6] into the future and solve for the time t_{vac} at which the universe becomes vacuum-dominated. We find the result

$$t_{\rm vac} = t_0 + \tau \frac{\sinh^{-1}\left[1\right] - \sinh^{-1}\left[\nu^{1/2}\right]}{\nu^{1/2}}, \tag{5.35}$$

where t_0 is the present age of the universe and we have defined $\tau \equiv (6\pi G\rho_0)^{-1/2}$; both time scales t_0 and τ are approximately 10^{10} yr.

Several results are immediately apparent from equation [5.35]. If the vacuum energy density provides any appreciable fraction of the total energy density at the present epoch (in other words, if ν is not too small), then the universe will enter an inflationary phase in the very near future. Furthermore, almost any nonvanishing value of the present day vacuum energy will lead the universe into an inflationary phase on the long time scales considered in this paper. For small values of the ratio ν, the future inflationary epoch occurs at the cosmological decade given by

$$\eta_{\rm inflate} \approx 10 + \frac{1}{2}\log_{10}\left[\frac{1}{\nu}\right]. \tag{5.36}$$

For example, even for a present-day vacuum contribution as small as $\nu \sim 10^{-40}$, the universe will enter an inflationary phase at the cosmological decade $\eta_{inflate} \approx 30$, long before protons begin to decay. In other words, the traditional cosmological constant problem becomes even more severe when we consider future cosmological decades.

If the universe enters into a future inflationary epoch, several interesting consequences arise. After a transition time comparable to the age of the universe at the epoch [5.36], the scale factor of the universe will begin to grow superluminally. Because of this rapid expansion, all of the astrophysical objects in the universe become isolated and eventually become out of causal contact. In other words, every given co-moving observer will see an effectively *shrinking horizon* (the particle horizon does not actually get smaller, but this language has become common in cosmology – see Ellis & Rothman, 1993 for further discussion of horizons in this context). In particular, astrophysical objects, such as galaxies and stars, will cross outside the speed-of-light sphere and hence disappear from view. For these same astrophysical objects, the velocity relative to the observer becomes larger than the speed of light and their emitted photons are redshifted to infinity.

2 *Tunneling processes*

We next consider the possibility that the universe is currently in a false vacuum state. In other words, a lower-energy vacuum state exists and the universe can someday tunnel to that lower-energy state. This problem, the fate of the false vacuum, was first explored quantitatively by Voloshin et al. (1974) and by Coleman (1977). Additional effects have been studied subsequently, including gravity (Coleman & De Luccia, 1980) and finite-temperature effects (e.g., Linde, 1983).

To obtain quantitative results, we consider an illustrative example in which the vacuum energy density of the universe can be described by the dynamics of a single scalar field. Once a field configuration becomes trapped in a metastable state (the false

vacuum), bubbles of the true vacuum state nucleate in the sea of false vacuum and begin growing spherically. The speed of the bubble walls quickly approaches the speed of light. The basic problem is to calculate the tunneling rate (the decay probability) from the false vacuum state to the true vacuum state, i.e., the bubble nucleation rate P per unit time per unit volume. For tunneling of scalar fields at zero temperature (generally called quantum tunneling), the four-dimensional Euclidean action S_4 of the theory largely determines this tunneling rate. The decay probability P can be written in the form

$$P = K\,\mathrm{e}^{-S_4}\,, \tag{5.37}$$

where K is a determinental factor (see Coleman, 1977, 1985). For purposes of illustration, we assume a generic quartic potential of the form

$$V(\Phi) = \lambda\Phi^4 - a\Phi^3 + b\Phi^2 + c\Phi + d \tag{5.38}$$

We can then write the action S_4 in the form

$$S_4 = \frac{\pi^2}{3\lambda}(2-\delta)^{-3} R(\delta)\,, \tag{5.39}$$

where $\delta \equiv 8\lambda b/a^2$ and where R is a slowly varying function which has a value near unity for most of the range of possible quartic potentials (Adams, 1993). The composite shape parameter δ varies from 0 to 2 as the potential $V(\Phi)$ varies from having no barrier height to having nearly degenerate vacua (see Figure 5.10).

Even though equations [5.37 – 5.39] describe the tunneling rate, we unfortunately do not know what potential (if any) describes our universe and hence it is difficult to obtain a precise numerical estimate for this time scale. To get some quantitative feeling for this problem, we consider the following example. For the case of no tunneling barrier (i.e., for $S_4 = 0$), the characteristic decay probability is given by $P_0 \sim K \sim M_V^4$, where M_V is the characteristic energy scale for the scalar field. For $M_V = 10^{16}$ GeV (roughly the GUT scale), $P_0 \sim 10^{129}$ s^{-1} cm^{-3}. With this decay rate, the universe within a characteristic volume M_V^{-3} would convert from false vacuum to true vacuum on a time scale of $\sim 10^{-24}$ s. Clearly, however, the actual decay time scale must be long enough that the universe has not decayed by the present epoch. In order to ensure that the universe has survived, we require that no nucleation events have occurred within the present horizon volume (~[3000 Mpc]^3) during the current age of the universe (~10^{10} yr). This constraint implies that the action S_4 must be sufficiently large in order to suppress nucleation, in particular,

$$S_4 > 231 \ln 10 \approx 532\,. \tag{5.40}$$

The question then becomes: is this value for S_4 reasonable? For the parameter λ, a reasonable range of values is $0.1 < \lambda < 1$; similarly, for δ, we take the range $0.1 < \delta < 1.9$. Using the form [5.39] for the action and setting $R = 1$, we find the approximate range

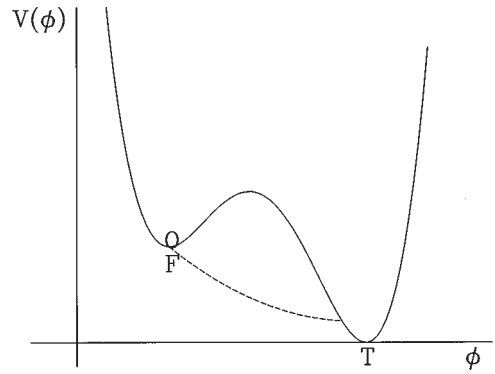

Figure 5.10 Potential V (Φ) of a scalar field which determines the vacuum state of the universe. This potential has both a false vacuum state (labeled F) and a true vacuum state (labeled T). As illustrated by the dashed curve, the universe can tunnel from the false vacuum state into the true vacuum state at some future time.

$$0.5 < S_4 < 3\times 10^4 \,. \tag{5.41}$$

Thus, the value required for the universe to survive to the present epoch (equation [5.40]) can be easily realized within this simple model. In the future, however, the universe could tunnel into its false vacuum state at virtually any time, as soon as tomorrow, or as late as $\eta = 10^4$. If and when this tunneling effect occurs, the universe will change its character almost completely. The physical laws of the universe, or at least the values of all of the physical constants, would change as the phase transition completes (see Sher, 1989 and Crone & Sher, 1990 for a discussion of changing laws of physics during a future phase transition). The universe, as we know it, would simply cease to exist.

Vacuum tunneling of the entire universe is certainly one of the more speculative topics considered in this paper. Nevertheless, its inclusion is appropriate since the act of tunneling from a false vacuum into a true vacuum would change the nature of the universe more dramatically than just about any other physical process.

It is also possible for the universe to spontaneously create "child universes" through a quantum tunneling process roughly analogous to that considered above (e.g., Sato et al., 1982; Hawking, 1987; Blau, Guendelman, & Guth, 1987). In this situation, a bubble of false vacuum energy nucleates in an otherwise empty space-time. If this bubble is sufficiently large, it will grow exponentially and will eventually become causally disconnected from the original space-time. In this sense, the newly created bubble becomes a separate "child universe". The newly created universe appears quite different to observers inside and outside the bubble. Observers inside the bubble see the local universe in a state of exponential expansion. Observers outside the bubble, in the empty space-time background, see the newly created universe as a black hole that collapses and becomes causally disconnected. As a result, these child universes will

not greatly affect the future evolution of our universe because they (relatively) quickly become out of causal contact.

One potentially interesting effect of these child universes is that they can, in principle, receive information from our universe. Before the newly created universe grows out of causal contact with our own universe, it is connected through a relativistic wormhole, which can provide a conduit for information transfer and perhaps even the transfer of matter (see Visser, 1995 for further discussion of wormholes and transferability). The implications of this possibility are the subject of current debate (for varying points of view, see, e.g., Linde, 1988, 1989; Tipler, 1992; Davies, 1994).

F Speculations about energy and entropy production in the far future

Thus far in this paper, we have shown that entropy can be generated (and hence work can be done) up to cosmological decades $\eta \sim 100$. For very long time scales $\eta \gg 100$, the future evolution of the universe becomes highly uncertain, but the possibility of continued entropy production is very important (see §VI.D). Here, we briefly assess some of the possible ways for energy and entropy to be generated in the far future.

1 Continued formation and decay of black holes

For the case of a flat spatial geometry for the universe, future density perturbations can provide a mechanism to produce entropy. These density perturbations create large structures which can eventually collapse to form black holes. The resulting black holes, in turn, evaporate by emitting Hawking radiation and thus represent entropy (and energy) sources (e.g., see also Page & McKee, 1981a; Frautschi, 1982). Density perturbations of increasingly larger size scale λ will enter the horizon as the universe continues to expand. The corresponding mass scale M_λ of these perturbations is given by

$$M_\lambda = M_0 \left(\frac{t_\lambda}{t_0} \right). \tag{5.42}$$

where t_λ is the time at which the perturbation enters the horizon and $M_0 \approx 10^{22} M_e$ is the total mass within the present-day horizon (at time t_0).

The time t_λ represents the time at which a given perturbation enters the horizon and begins to grow; a large structure (such as a black hole) can only form at some later time after the perturbation becomes nonlinear. Suppose that a density perturbation has an initial amplitude δ_λ when it enters the horizon. In the linear regime, the perturbation will grow according to the usual relation

$$\delta = \delta_\lambda \left(\frac{t}{t_\lambda} \right)^{2/3}, \tag{5.43}$$

where $\delta \equiv \Delta\rho/\rho$ and $t > t_\lambda$ (see Peebles, 1993). Using this growth law, the epoch η_{nl} at which the perturbation becomes nonlinear can be written in the form

$$\eta_{\rm nl} = \eta_\lambda - \frac{3}{2}\log_{10}\delta_\lambda \ . \tag{5.44}$$

For example, if the perturbation has an amplitude $\delta_\lambda = 10^{-4}$, then it becomes nonlinear at time $\eta_{nl} = \eta_\lambda + 6$. Since we are interested in very long time scales $\eta > 100$, the difference between the horizon crossing time η_λ and the time η_{nl} of nonlinearity is not overly large.

One possible result of this process is the production of a large black hole with a mass $M_{BH} \sim M_\lambda$. The time scale for such a black hole to evaporate through the Hawking process is given by

$$\eta_{BH} = 101 + 3\eta_\lambda \ , \tag{5.45}$$

where we have combined equations [4.42] and [5.42]. Since $\eta_{BH} \gg \eta_\lambda \sim \eta_{nl}$, the universe can form black holes faster than they can evaporate. Thus, for the case of a geometrically flat universe, future density perturbations can, in principle, continue to produce black holes of increasingly larger mass. In this case, the universe will always have a source of entropy – the Hawking radiation from these black holes.

We note that these bound perturbations need not necessarily form black holes. The material is (most likely) almost entirely non-dissipative and collisionless, and will thus have a tendency to form virialized clumps with binding energy per unit mass of order $\sim \delta c^2$. Thus, unless the perturbation spectrum is tilted so that δ is of order unity on these much larger scales, the ensuing dynamics is probably roughly analogous to that of a cluster-mass clump of cold dark matter in our present universe. However, even if the mass of the entire perturbation does not form a single large black hole, smaller scale structures can in principle form black holes, in analogy to those currently in the centers of present-day galaxies. In addition, it is possible that the existing black holes can merge faster than they evaporate through the Hawking process (see also §III.D). Thus, the possibility remains for the continued existence of black holes in the universe.

The process outlined here, the formation of larger and larger black holes, can continue as long as the universe remains spatially flat and the density perturbations that enter the horizon are not overly large. The inflationary universe scenario provides a mechanism to achieve this state of affairs, at least up to some future epoch (see §V.C and in particular equation [5.14]). Thus, the nature of the universe in the far future $\eta \gg 100$ may be determined by the physics of the early universe (in particular, inflation) at the cosmological decade $\eta \sim -45$.

Notice that at these very late times, $\eta \gg 100$, the matter entering the horizon will already be "processed" by the physical mechanisms described earlier in the paper. Thus, the nucleons will have (most likely) already decayed and the matter content of the universe will be mostly electrons, positrons, and non-baryonic dark matter particles. Annihilation of both e^+–e^- pairs and dark matter will occur simultaneously

with perturbation growth and hence the final mass of the black hole will be less than M_λ. This issue must be studied in further depth.

2 Particle annihilation in an open universe

If the universe is open, however, then future density perturbations are effectively frozen out (see §V.B) and the hierarchy of black holes described above cannot be produced. For an open universe, continued energy and entropy production is more difficult to achieve. One process that can continue far into the future, albeit at a very low level, is the continued annihilation of particles. Electrons and positrons represent one type of particle that can annihilate (see also Page & McKee, 1981ab), but the discussion given below applies to a general population of particles.

Consider a collection of particles with number density *n*. The time evolution of the particle population is governed by the simple differential equation

$$\frac{dn}{dt} + 3H_n = -\langle \sigma \upsilon \rangle n^2 , \tag{5.46}$$

where $H = \dot{R}/R$ is the Hubble parameter and $\langle \sigma v \rangle$ is the appropriate average of interaction cross section times the speed (e.g., see Kolb & Turner, 1990). Since we are interested in the case for which the expansion rate is much larger than the interaction rate, the particles are very far from thermal equilibrium and we can neglect any back reactions that produce particles. For this example, we consider the universe to be open, independent of the activity of this particle population. As a result, we can write $R \propto t$ and hence $H = 1/t$. We also take the quantity $\langle \sigma v \rangle$ to be a constant in time (corresponding to *s*-wave annihilation). With these approximations, the differential equation [5.46] can be integrated to obtain the solution

$$n(t) = n_1 \left(\frac{t_1}{t} \right)^3 \left[1 + \Delta_\infty \left[1 - \left(t_1 / t \right)^2 \right] \right]^{-1} , \tag{5.47}$$

where we have defined the quantity

$$\Delta_\infty \equiv \frac{1}{2} n_1 t_1 \langle \sigma \upsilon \rangle , \tag{5.48}$$

and where we have invoked the boundary condition

$$n(t_1) = n_1 = constant . \tag{5.49}$$

Analogous solutions for particle annihilation can be found for the case of a flat universe ($H = 2/3t$) and an inflating universe ($H = constant$).

The difference between the solution [5.47] and the simple adiabatic scaling solution $n(t) = n_1(t_1/t)^3$ is due to particle annihilation, which is extremely small but non-zero.

This statement can be quantified by defining the fractional difference Δ between the solution [5.47] and the adiabatic solution, i.e.,

$$\Delta(t) \equiv \frac{\Delta n}{n}(t) = \Delta_{\infty}\left[1-\left(t_{1}/t\right)^{2}\right]. \tag{5.50}$$

Over the entire (future) lifetime of the universe, the comoving fraction of particles that annihilate is given by the quantity Δ_{∞}, which is both finite and typically much less than unity. For example, if we consider the largest possible values at the present epoch ($\sigma \approx \sigma_{T} \approx 10^{-24}$ cm^{2}, $n_{1} \approx 10^{-6}$ cm^{-3}, $t_{1} \approx 3 \times 10^{17}$ *s*, and $v = c$), then $\Delta_{\infty} \approx 10^{-2}$. The fraction Δ_{∞} will generally be much smaller than this example. The fact that the fraction Δ_{∞} is finite implies that the process of particle annihilation can provide only a finite amount of energy over the infinite time interval $\eta_{1} < \eta < \infty$.

3 Formation and decay of positronium

Another related process that will occur on long time scales is the formation and eventual decay of positronium. This process has been studied in some detail by Page & McKee (1981ab; see also the discussion of Barrow & Tipler, 1986); here we briefly summarize their results. The time scale for the formation of positronium in a flat universe is given by

$$\eta_{\text{form}} \approx 85 + 2\left(\eta_{P} - 37\right) - \frac{2}{3}\log_{10}\left[\Omega_{e}\right]. \tag{5.51}$$

where η_{P} is the proton lifetime (see §IV) and where Ω_{e} is the mass fraction of $e^{\pm}$ after proton decay. For a flat or nearly flat universe, most of the electrons and positrons become bound into positronium. In an open universe, some positronium formation occurs, but most electrons and positrons remain unattached.

At the time of formation, the positronium atoms are generally in states of very high quantum number (and have radii larger than the current horizon size). The atoms emit a cascade of low-energy photons until they reach their ground state; once this occurs, the positronium rapidly annihilates. The relevant time scale for this decay process is estimated to be

$$\eta_{\text{decay}} \approx 141 + 4\left(\eta_{P} - 37\right) - \frac{8}{3}\log_{10}\left[\Omega_{e}\right]. \tag{5.52}$$

VI Summary and discussion

Our goal has been to present a plausible and quantitative description of the future of the Universe. Table 5.1 outlines the most important events in the overall flow of time, as well as the cosmological decades at which they occur (see equation [1.1]). In constructing this table, representative values for the (often uncertain)

Table 5.1 Important Events in the History and Future of the Universe

The Big Bang	$\eta = -\infty$
Planck Epoch	−50.5
GUT Epoch	−44.5
Electroweak phase transition	−17.5
Quarks become confined into hadrons	−12.5
Nucleosynthesis	−6
* * * * * * * * * *	
Matter domination	4
Recombination	5.5
First possible stellar generation	6
Formation of the Galaxy	9
Formation of the Solar System	9.5
Today: The Present Epoch	10
Our Sun dies	10.2
Close Encounter of Milky Way with Andromeda (M31)	10.2
Lower bound on the age of closed universe	10.8
Lifetime of main-sequence stars with lowest mass	13
End of conventional star formation	14
* * * * * * * * * *	
Planets become detached from stars	15
Star formation via brown dwarf collisions	16
Lower bound on age of flat universe (with future $\Delta\rho/\rho > 1$)	18
Stars evaporate from the Galaxy	19
Planetary orbits decay via gravitational radiation	20
WIMPs in the galactic halo annihilate	22.5
Star formation via orbital decay of brown dwarf binaries	23
Stellar orbits in the galaxy decay via gravitational radiation	24
White dwarfs deplete WIMPs from the galactic halo	25
Black holes accrete stars on galactic size scale	30
Black holes accrete stars on cluster size scale	33
Protons decay	37
Neutron stars β-decay	38
Planets destroyed by proton decay	38

White Dwarfs destroyed by proton decay	39
* * * * * * * * * *	
Axions decay into photons	42
Hydrogen molecules experience pycnonuclear reactions	60
Stellar-sized black holes evaporate	65
Black holes with $M = 10^6 M_e$ evaporate	83
Positronium formation in a flat universe	85
Galaxy-sized black holes evaporate	98
Black hole with mass of current horizon scale evaporates	131
Positronium decay in a flat universe	141
Higher-order proton decay processes	~100 – 200
* * * * * * * * * *	

parameters have been assumed; the stated time scales must therefore be viewed as approximate. Furthermore, as a general rule, both the overall future of the universe, as well as the time line suggested in Table 5.1 become more and more uncertain in the face of successively deeper extrapolations into time. Some of the effects we have described will compete with one another, and hence not all the relevant physical processes can proceed to completion. Almost certainly, parts of our current time line will undergo dramatic revision as physical understanding improves. We have been struck by the remarkable natural utility of the logarithmic "clock", η, in organizing the passage of time. Global processes which can characterize the entire universe rarely span more than a few cosmological decades, and the ebb and flow of events is dispersed quite evenly across a hundred and fifty orders of magnitude in time, i.e., $-50 < \eta < 100$.

A Summary of results

Our specific contributions to physical eschatology can be summarized as follows:

[1] We have presented new stellar evolution calculations which show the long-term behavior of very low mass stars (see Figure 5.1). Stars with very small-mass ($\sim 0.1 M_e$) do not experience any red giant phases. As they evolve, these stars become steadily brighter and bluer, reaching first a maximum luminosity, and second, a maximum temperature, prior to fading away as helium white dwarfs.

[2] Both stellar evolution and conventional star formation come to an end at the cosmological decade $\eta \sim 14$. This time scale only slightly exceeds the longest evolution time for a low-mass star. It also corresponds to the time at which the galaxy runs out of raw material (gas) for producing new stars. The era of conventional stars in the universe is confined to the range $6 < \eta < 14$.

[3] We have introduced the final mass function (FMF), i.e., the distribution of masses for the degenerate stellar objects left over from stellar evolution (see Figure 5.2). Roughly half of these objects will be white dwarfs, with most of the remainder being brown dwarfs. Most of the mass, however, will be in the form of white dwarfs (see equations [2.22] and [2.23]).

[4] We have explored a new mode of continued star formation through the collisions of substellar objects (see Figure 5.3). Although the time scale for this process is quite long, this mode of star formation will be the leading source of new stars for cosmological decades in the range $15 < \eta < 23$.

[5] We have presented a scenario for the future evolution of the galaxy. The galaxy lives in its present state until a time of $\eta \sim 14$ when both conventional star formation ceases and the smallest ordinary stars leave the main sequence. For times $\eta > 14$, the principal mode of additional star formation is through the collisions and mergers of brown dwarfs (substellar objects). The galaxy itself evolves through the competing processes of orbital decay of orbits via gravitational radiation and the evaporation of stars into the intergalactic medium via stellar encounters. Stellar evaporation is the dominant process and most of the stars will leave the system at a time $\eta \sim 19$. Some fraction (we roughly estimate ~0.01–0.10) of the galaxy is left behind in its central black hole.

[6] We have considered the annihilation and capture of weakly interacting massive particles (WIMPs) in the galactic halo. In the absence of other evolutionary processes, the WIMPs in the halo annihilate on the time scale $\eta \sim 23$. On the other hand, white dwarfs can capture WIMPs and thereby deplete the halo on the somewhat longer time scale $\eta \sim 25$. The phenomenon of WIMP capture indicates that white dwarf cooling will be arrested rather shortly at a luminosity $L_* \sim 10^{-12} L_e$.

[7] Depending on the amount of mass-loss suffered by the Sun when it becomes a red giant, the Earth may be vaporized by the Sun during its asymptotic giant phase of evolution; in this case, the Earth will be converted to a small (0.01 %) increase in the solar metallicity. In general, however, planets can end their lives in a variety of ways. They can be vaporized by their parent stars, ejected into interstellar space through close stellar encounters, merge with their parent stars through gravitational radiation, and can eventually disappear as their protons decay.

[8] We have discussed the allowed range for the proton lifetime. A firm lower bound on the lifetime arises from current experimental searches. Although no definitive upper limit exists, we can obtain a suggestive upper "bound" on the proton lifetime by using decay rates suggested by GUTs and by invoking the constraint the mass of the mediating boson, $M_X < M_{P1} \sim 10^{19}$ GeV. We thus obtain the following expected range for the proton lifetime

$$32 < \eta_P < 49 + 76\left(N - 1\right), \tag{6.1}$$

where the integer N is order of the process, i.e., the number of mediating bosons required for the decay to take place. Even for the third order case, we have $\eta_P <$

201. Quantum gravity effects also lead to proton decay with time scales in the range $46 < \eta_P < 169$. Finally, sphalerons imply $\eta_P \sim 140$.

[9] We have presented a scenario for the future evolution of sun-like stars (see Figure 5.6). In this case, stars evolve into white dwarf configurations as in conventional stellar evolution. On sufficiently long time scales, however, proton decay becomes important. For cosmological decades in the range $20 < \eta < 35$, the mass of the star does not change appreciably, but the luminosity is dominated by the energy generated by proton decay. In the following cosmological decades, $\eta = 35 - 37$, mass-loss plays a large role in determining the stellar structure. The star expands as it loses mass and follows the usual mass/radius relation for white dwarfs. The chemical composition changes as well (see Figure 5.5). Proton decay by itself quickly reduces the star to a state of pure hydrogen. However, pycnonuclear reactions will be sufficient to maintain substantial amounts of helium (^{3}He and ^{4}He) until the mass of the star decreases below $\sim 0.01 M_e$. During the proton-decay phase of evolution, a white dwarf follows a well-defined track in the H-R Diagram given by $L_\star \propto T_\star^{12/5}$. After the stellar mass decreases to $M_\star \approx 10^{-3} M_e$, the star is lifted out of degeneracy and follows a steeper track $L_\star \propto T_\star^{12}$ in the H-R Diagram.

[10] If proton decay does not take place through the first-order process assumed above, then white dwarfs and other degenerate objects will still evolve, but on a much longer time scale. The relevant physical process is likely to be proton decay through higher-order effects. The time scales for the destruction and decay of degenerate stars obey the ordering

$$\eta_P \ll \eta_{BH} \ll \eta_{P2}\,, \tag{6.2}$$

where $\eta_P \sim 37$ is the time scale for first-order proton decay, $\eta_{BH} \sim 65$ is the time scale for a stellar-sized black hole to evaporate, and $\eta_{P2} \sim 100 - 200$ is the time scale for proton decay through higher-order processes.

[11] In the future, the universe as a whole can evolve in a variety of different possible ways. Future density perturbations can come across the horizon and close the universe; this effect would ultimately lead (locally) to a big crunch. Alternately, the universe could contain a small amount of vacuum energy (a cosmological-constant term) and could enter a late time inflationary epoch. Finally, the universe could be currently in a false vacuum state and hence kevorking on the brink of instability. In this case, when the universe eventually tunnels into the true vacuum state, the laws of physics and hence the universe as we know it would change completely.

[12] As the cosmic microwave background redshifts away, several different radiation fields will dominate the background. In the near term, stellar radiation will overtake the cosmic background. Later on, the radiation produced by dark matter annihilation (both direct and in white dwarfs) will provide the dominant contribution. This radiation field will be replaced by that arising from proton decay, and then, eventually, by the radiation field arising from evaporation of black holes (see Figure 5.9).

B Eras of the future universe

Our current understanding of the universe suggests that we can organize the future into distinct eras, somewhat analogous to geological eras:

[A] *The Radiation-Dominated Era.* $-\infty < \eta < 4$. This era corresponds to the usual time period in which most of the energy density of the universe is in the form of radiation.

[B] *The Stelliferous Era.* $6 < \eta < 14$. Most of the energy generated in the universe arises from nuclear processes in conventional stellar evolution.

[C] *The Degenerate Era.* $15 < \eta < 37$. Most of the (baryonic) mass in the universe is locked up in degenerate stellar objects: brown dwarfs, white dwarfs, and neutron stars. Energy is generated through proton decay and particle annihilation.

[D] *The Black Hole Era.* $38 < \eta < 100$. After the epoch of proton decay, the only stellar-like objects remaining are black holes of widely disparate masses, which are actively evaporating during this era.

[E] *The Dark Era.* $\eta > 100$. At this late time, protons have decayed and black holes have evaporated. Only the waste products from these processes remain: mostly photons of colossal wavelength, neutrinos, electrons, and positrons. The seeming poverty of this distant epoch is perhaps more due to the difficulties inherent in extrapolating far enough into the future, rather than an actual dearth of physical processes.

C Experimental and theoretical implications

Almost by definition, direct experiments that test theoretical predictions of the very long-term fate of the universe cannot be made in our lifetimes. However, this topic in general and this paper in particular have interesting implications for present day experimental and theoretical work. If we want to gain more certainty regarding the future of the universe and the astrophysical objects within it, then several issues must be resolved. The most important of these are as follows:

[A] Does the proton decay? What is the lifetime? This issue largely determines the fate stellar objects in the universe for time scales longer than $\eta \sim 35$. If the proton is stable to first-order decay processes, then stellar objects in general and white dwarfs in particular can live in the range of cosmological decades $\eta < 100$. If the proton is also stable to second-order decay processes, then degenerate stellar objects can live for a much longer time. On the other hand, if the proton does decay, a large fraction of the universe will be in the form of proton decay products (neutrinos, photons, positrons, etc.) for times $\eta > 35$.

[B] What is the vacuum state of the universe? This issue plays an important role in determining the ultimate fate of the universe itself. If the vacuum energy density of the universe is nonzero, then the universe might ultimately experience a

future epoch of inflation. On the other hand, if the vacuum energy density is strictly zero, then future (large) densities perturbations can, in principle, enter our horizon and lead (locally) to a closed universe and hence a big crunch.

[C] What is the nature of the dark matter? Of particular importance is the nature of the dark matter that makes up galactic halos. The lifetime of the dark matter particles is also of great interest.

[D] What fraction of the stars in a galaxy are evaporated out of the system and what fraction are accreted by the central black hole (or black holes)? This issue is important because black holes dominate the energy and entropy production in the universe in the time range $36 < \eta < 100$ and the mass of a black hole determines its lifetime.

[E] Does new physics occur at extremely low temperatures? As the universe evolves and continues to expand, the relevant temperatures become increasingly small. In the scenario outlined here, photons from the cosmic microwave background and other radiation fields, which permeate all of space, can redshift indefinitely in accordance with the classical theory of radiation. It seems possible that classical theory will break down at some point. For example, in an open universe, the CMB photons will have a wavelength longer than the current horizon size ($\sim$ 3000 Mpc) at a time $\eta \sim 40$, just after proton decay. Some preliminary models for future phase transitions have been proposed (Primack & Sher, 1980; Suzuki, 1988; Sher, 1989), but this issue calls out for further exploration.

D Entropy and heat death

The concept of the heat death of the universe has troubled many philosophers and scientists since the mid-nineteenth century when the second law of thermodynamics was first understood (e.g., Helmholz, 1854; Clausius, 1865, 1868). Very roughly, *classical heat death* occurs when the universe as a whole reaches thermodynamic equilibrium; in such a state, the entire universe has a constant temperature at all points in space and hence no heat engine can operate. Without the ability to do physical work, the universe "runs down" and becomes a rather lifeless place. Within the context of modern Big Bang cosmology, however, the temperature of the universe is continually changing and the issue shifts substantially; many authors have grappled with this problem, from the inception of Big Bang theory (e.g., Eddington, 1931) to more recent times (Barrow & Tipler, 1978, 1986; Frautschi, 1982). A continually expanding universe never reaches true thermodynamic equilibrium and hence never reaches a constant temperature. Classical heat death is thus manifestly avoided. However, the expansion can, in principle, become purely adiabatic so that the entropy in a given comoving volume of the universe approaches (or attains) a constant value. In this case, the universe can still become a dull and lifeless place with no ability to do physical work. We denote this latter possibility as *cosmological heat death*.

Long-term entropy production in the universe is constrained in fairly general terms for a given class of systems (Bekenstein, 1981). For a spatially bounded physical system with effective radius *R,* the entropy *S* of the system has a well defined maximum value. This upper bound is given by

$$S \leq \frac{2\pi RE}{\hbar c}, \tag{6.3}$$

where *E* is the total energy of the system. Thus, for a bounded system (with finite size *R*), the ratio *S/E* of entropy to energy has a firm upper bound. Furthermore, this bound can be actually attained for black holes (see Bekenstein, 1981 for further discussion).

The results of this paper show that cosmological events continue to produce energy and entropy in the universe, at least until the cosmological decade $\eta \sim 100$. As a result, cosmological heat death is postponed until after that epoch, i.e., until the Dark Era. After that time, however, it remains possible in principle for the universe to become nearly adiabatic and hence dull and lifeless. The energy- and entropy-generating mechanisms available to the universe depend on the mode of long-term evolution, as we discuss below.

If the universe is closed (§V.A) or becomes closed at some future time (§V.B), then the universe will end in a big crunch and long-term entropy production will not be an issue. For the case in which the universe remains nearly flat, density perturbations of larger and larger size scales can enter the horizon, grow to nonlinearity, and lead to continued production of energy and entropy through the evaporation of black holes (see §V.F.1). These black holes saturate the Bekenstein bound and maximize entropy production. Cosmological heat death can thus be avoided as long as the universe remains nearly flat.

On the other hand, if the universe is open, then density fluctuations become frozen out at some finite length scale (§V.B). The energy contained within the horizon thus becomes a finite quantity. However, the Bekenstein bound does not directly constrain entropy production in this case because the effective size *R* grows without limit. For an open universe, the question of cosmological heat death thus remains open. For a universe experiencing a future inflationary phase (§V.E.1), the situation is similar. Here, the horizon is effectively shrinking with time. However, perturbations that have grown to nonlinearity will be decoupled from the Hubble flow. The largest nonlinear perturbation will thus define a largest length scale λ and hence a largest mass scale in the universe; this mass scale once again implies a (finite) maximum possible amount of energy available to a local region of space. However, the system is not bounded spatially and the questions of entropy production and cosmological heat death again remain open.

To close this paper, we put forth the point of view that the universe should obey a type of *Copernican Time Principle* which applies to considerations of the future. This principle holds that the current cosmological epoch ($\eta = 10$) has no special place in time. In other words, interesting things can continue to happen at the increasingly low levels of energy and entropy available in the universe of the future.

Acknowledgments

This paper grew out of a special course taught at the University of Michigan for the theme semester "Death, Extinction, and the Future of Humanity" (Winter 1996). We would like to thank Roy Rappaport for providing the initial stimulation for this course and hence this paper. We also thank R. Akhoury, M. Einhorn, T. Gherghetta, G. Kane, and E. Yao for useful discussions regarding proton decay and other particle physics issues. We thank P. Bodenheimer, G. Evrard, J. Jijina, J. Mohr, M. Rees, D. Spergel, F. X. Timmes, and R. Watkins for many interesting astrophysical discussions and for critical commentary on the manuscript. This work was supported by an NSF Young Investigator Award, NASA Grant No. NAG 5-2869, and by funds from the Physics Department at the University of Michigan.

References

Abramowitz, M., and I. A. Stegun, 1972, *Handbook of Mathematical Functions* (New York: Dover).
Adams, F. C. 1993, Phys. Rev. D **48**, 2800.
Adams, F. C., and M. Fatuzzo, 1996, Astrophys. J. **464**, 256.
Adams, F. C., and K. Freese, 1991, Phys. Rev. D **43**, 353.
Adams, F. C., and K. Freese, 1995, Phys. Rev. D **51**, 6722.
Adams, F. C., K. Freese, and A. H. Guth, 1991, Phys. Rev. D **43**, 965.
Albrecht, A., and P. J. Steinhardt, 1982, Phys. Rev. Lett. **48**, 1220.
Alcock, C. et al. 1993, Nature **365**, 621.
Aubourg, E. et al. 1993, Nature **365**, 623.
Bahcall, J. N. 1989, *Neutrino Astrophysics* (Cambridge: Cambridge Univ. Press).
Bardeen, J. M., P. J. Steinhardt, and M. S. Turner, 1983, Phys. Rev. D **28**, 679.
Barrow, J. D., and F. J. Tipler, 1978, Nature **276**, 453.
Barrow, J. D., and F. J. Tipler, 1986, *The Anthropic Cosmological Principle* (Oxford: Oxford Univ. Press).
Bekenstein, J. D. 1981, Phys. Rev. D **23**, 287.
Binney, J., and S. Tremaine, 1987, *Galactic Dynamics* (Princeton: Princeton Univ. Press).
Blau, S. K., E. I. Guendelman, and A. H. Guth 1987, Phys. Rev. D **35**, 1747.
Laughlin, G., P. Bodenheimer, and F. C. Adams, 1996, submitted to Astrophys. J.
Bond, J. R., B. J. Carr, and C. J. Hogan, 1991, Astrophys. J. **367**, 420.
Brune, D., and J. J. Schmidt, 1974, editors, *Handbook on Nuclear Activation Cross-Sections* (Vienna: International Atomic Energy Agency).
Burrows, A., W. B. Hubbard, D. Saumon, and J. I. Lunine, 1993, Astrophys. J. **406**, 158.
Burrows, A., and J. Liebert, 1993, Rev. Mod. Phys. 65, 301.
Carroll, S. M., W. H. Press, and E. L. Turner, 1992, Ann. Rev. Astron. Astrophys. **30**, 499.
Castano, D. J., and S. P. Martin, 1994, Phys. Lett. 340 B, 67.
Chandrasekhar, S. 1939, *Stellar Structure* (New York: Dover).
Clausius, R. 1865, Ann. Physik, **125**, 353.
Clausius, R. 1868, Phil. Mag., **35**, 405.

Clayton, D. D. 1983, *Principles of Stellar Evolution and Nucleosynthesis* (Chicago: Univ. Chicago Press).
Coleman, S. 1977, Phys. Rev. D **15**, 2929.
Coleman, S. 1985, *Aspects of Symmetry* (Cambridge, England: Cambridge Univ. Press).
Coleman, S., and F. De Luccia, 1980, Phys. Rev. D **21**, 3305.
Copeland, H., J. O. Jensen, and H. E. Jorgensen, 1970, Astron. Astrophys. **5**, 12.
Crone, M. M., and M. Sher, 1990, Am. J. Phys. **59**, 25.
D'Antona, F., and I. Mazzitelli, 1985, Astrophys. J. **296**, 502.
Davies, P.C.W. 1994, *The Last Three Minutes* (New York: BasicBooks).
Dicus, D. A., J. R. Letaw, D. C. Teplitz, and V. L. Teplitz, 1982, Astrophys. J. **252**, 1.
Diehl, E., G. L. Kane, C. Kolda, and J. D. Wells, 1995, Phys. Rev. D **52**, 4223.
Dorman, B., L. A. Nelson, and W. Y. Chan, 1989, Astrophys. J. **342**, 1003.
Draine, B. T., and H. M. Lee, 1984, Astrophys. J. **285**, 89.
Dyson, F. J. 1979, Rev. Mod. Phys. **51**, 447.
Dyson, F. J. 1988, *Infinite in All Directions* (New York: Harper and Row).
Eddington, A. S. 1931, Nature **127**, 447.
Ellis, G.F.R., and T. Rothman, 1993, Am. J. Phys. **61**, 883.
Ellis, G.F.R., and D. H. Coule, 1994, Gen. Rel. and Grav. **26**, 731.
Elmegreen, B. G., and R. D. Mathieu, 1983, Mon. Not. R. Astron. Soc. **203**, 305.
Faulkner, J., and R. L. Gilliland, 1985, Astrophys. J. **299**, 994.
Feinberg, G., 1981, Phys. Rev. D **23**, 3075.
Feinberg, G., M. Goldhaber, and G. Steigman, 1978, Phys. Rev. D **18**, 1602.
Frautschi, S. 1982, Science **217**, 593.
Freese, K. 1986, Phys. Lett. 167 B, 295.
Gaier, T., et al. 1992, Astrophys. J. Lett. **398**, L1.
Goity, J. L., and M. Sher, 1995, Phys. Lett. 346 B, 69.
Golimowski, D. A., T. Nakajima, S. R. Kulkarni, and B. R. Oppenheimer, 1995, Astrophys. J. Lett. 444, L101.
Gott, J. R. III 1993, Nature **363**, 315.
Gould, A. 1987, Astrophys. J. **321**, 571.
Gould, A. 1991, Astrophys. J. **388**, 338.
Grischuk, L. P., and Ya. B. Zel'dovich, 1978, Sov. Astron. **22**, 125.
Grossman, A. S., and H. C. Graboske, 1971, Astrophys. J. **164**, 475.
Guth, A. 1981, Phys. Rev. D **23**, 347.
Guth, A. H., and S.-Y. Pi, 1982, Phys. Rev. Lett. **49**, 1110.
Hamada, T., and E. E. Salpeter, 1961, Astrophys. J. **134**, 683.
Hawking, S. W. 1975, Comm. Math. Phys. **43**, 199.
Hawking, S. W. 1982, Phys. Lett. 115 B, 295.
Hawking, S. W. 1985, Phys. Lett. 150 B, 339.
Hawking, S. W. 1987, Phys. Lett. 195 B, 337.
Hawking, S. W., D. N. Page, and C. N. Pope, 1979, Phys. Lett. 86 B, 175.
Helmholz, H. von1854, *On the Interaction of Natural Forces.*
Henry, T. J., J. D. Kirkpatrick, and D. A. Simons, 1994, Astron. J. **108**, 1437.
't Hooft, G. 1976, Phys. Rev. Lett. **37**, 8.
Hubbell, J. H., H. A. Grimm, and I. Overbo, 1980, J. Phys. Chem. Ref. Data **9**, 1023.
Islam, J. N. 1977, Quart. J. R. Astron. Soc. **18**, 3.
Islam, J. N. 1979, Sky and Telescope **57**, 13.

Jungman, G., M. Kamionkowski, and K. Griest, 1996, Physics Reports in press.
Jura, M. 1986, Astrophys. J. **301**, 624.
Kane, G. L. 1993, *Modern Elementary Particle Physics* (Reading MA: Addison–Wesley).
Kane, G. L., and J. D. Wells, 1996, Phys. Rev. Lett. in press.
Kennicutt, R. C., P. Tamblyn, and C. W. Congdon, 1995, Astrophys. J. **435**, 22.
Kippenhahn, R., and A. Weigert, 1990, *Stellar Structure and Evolution* (Berlin: Springer).
Kolb, E. W., and M. S. Turner, 1990, *The Early Universe* (Redwood City CA: Addison–Wesley).
Krauss, L. M., M. Srednicki, and F. Wilczek, 1986, Phys. Rev. D **33**, 2206.
Krauss, L. M., and M. White, 1992, Phys. Rev. Lett. **69**, 869.
Kumar, S. S. 1963, Astrophys. J. **137**, 1121.
La, D., and P. J. Steinhardt, 1989, Phys. Rev. Lett. **62**, 376.
Langacker, P. 1981, Physics Reports **72**, 186.
Larson, R. B. 1973, Mon. Not. R. Astron. Soc. **161**, 133.
Larson, R. B., and B. M. Tinsley, 1978, Astrophys. J. **219**, 46.
Laughlin, G., and P. Bodenheimer, 1993, Astrophys. J. **403**, 303.
Lee, H. M., and J. P. Ostriker, 1986, Astrophys. J. **310**, 176.
Lemaître, G. 1993, Ann. Soc. Sci. Bruxelles A53, 51.
Lightman, A. P., and S. L. Shapiro, 1978, Rev. Mod. Phys. **50**, 437.
Linde, A. D. 1982, Phys. Lett. 108 B, 389.
Linde, A. D. 1983, Nucl. Phys. B216, 421.
Linde, A. D. 1988, Phys. Lett. 211 B, 29.
Linde, A. D. 1989, Phys. Lett. 227 B, 352.
Linde, A. D. 1990, *Particle Physics and Inflationary Cosmology* (New York: Harwood Academic).
Loh, E., and E. Spillar, 1986, Astrophys. J. Lett. **307**, L1.
Lyth, D. H. 1984, Phys. Lett. 147 B, 403.
Manchester, R. N., and J. H. Taylor, 1977, *Pulsars* (San Francisco: W. H. Freeman).
Marcy, G. W., R. P. Butler, and E. Williams, 1996, submitted to Astrophys. J.
Mayor, M., and D. Queloz, 1995, Nature **378**, 355.
Meyer, S. S., E. S. Cheng, and L. A. Page, 1991, Astrophys. J. Lett. **410**, L57.
Mihalas, D., and J. Binney, 1981, *Galactic Astronomy: Structure and Kinematics* (New York: W. H. Freeman).
Miller, G. E., and J. M. Scalo, 1979, Astrophys. J. Suppl. **41**, 513.
Misner, C. W., K. S. Thorne, and J. A. Wheeler, 1973, *Gravitation* (San Francisco: W. H. Freeman).
Mohapatra, R. N., and R. E. Marshak, 1980, Phys. Rev. Lett. **44**, 1316.
Ohanian, H. C., and R. Ruffini, 1994, *Gravitation and Spacetime* (New York: W. W. Norton).
Oppenheimer, B. R., S. R. Kulkarni, K. Matthews, and T. Nakajima, 1995, Science **270**, 1478.
Page, D. N. 1980, Phys. Lett. 95 B, 244.
Page, D. N., and M. R. McKee, 1981a, Phys. Rev. D **24**, 1458.
Page, D. N., and M. R. McKee, 1981b, Nature **291**, 44.
Particle Data Group, 1994, Phys. Rev. D **50**, 1173.
Peccei, R. D., and H. R. Quinn 1977a, Phys. Rev. Lett. **38**, 1440.
Peccei, R. D., and H. R. Quinn 1977b, Phys. Rev. D **16**, 1791.

Peebles, P.J.E. 1993, *Principles of Physical Cosmology* (Princeton: Princeton Univ. Press).
Peebles, P.J.E. 1994, Astrophys. J. **429**, 43.
Perkins, D. 1984, Ann. Rev. Nucl. Part. Sci. **34**, 1.
Phillips, A. C. 1994, *The Physics of Stars* (Chichester: Wiley).
Poundstone, W. 1985, *The Recursive Universe* (New York: Morrow).
Press, W. H., and D. N. Spergel, 1985, Astrophys. J. **296**, 679.
Press, W. H., and S. A. Teukolsky, 1977, Astrophys. J. **213**, 183.
Primack, J. R., and M. Sher, 1980, Nature **288**, 680.
Rajaraman, R. 1987, *Solitons and Instantons* (Amsterdam: North-Holland).
Rana, N. C. 1991, Ann. Rev. Astron. Astrophys. **29**, 129.
Rees, M. J. 1969, Observatory **89**, 193.
Rees, M. J. 1981, Quart. J. R. Astron. Soc. **22**, 109.
Rees, M. J. 1984, Ann. Rev. Astron. Astrophys. **22**, 471.
Reimers, D. 1975, in *Problems in Stellar Astrophysics*, eds. B. Bascheck, W. H. Kegel, and G. Traving (New York: Springer), p. 229.
Riess, A. G., W. H. Press, and R. P. Kirshner, 1995, Astrophys. J. Lett. **438**, L17.
Roberts, M. S. 1963, Ann. Rev. Astron. Astrophys. **1**, 149.
Sackmann, I.-J., A. I. Boothroyd, and K. E. Kraemer, 1993, Astrophys. J. **418**, 457.
Sakharov, A. D. 1967, JETP Letters **5**, 24.
Salpeter, E. E. 1955, Astrophys. J. **121**, 161.
Salpeter, E. E. 1982, in *Essays in Nuclear Astrophysics* (Cambridge: Cambridge Univ. Press).
Salpeter, E. E., and H. M. Van Horn, 1969, Astrophys. J. **155**, 183.
Sato, K., H. Kodama, M. Sasaki, and K. Maeda, 1982, Phys. Lett. 108 B, 103.
Scalo, J. M. 1986, Fund. Cos. Phys. **11**, 1.
Schuster, J., et al. 1993, Astrophys. J. Lett. **412**, L47.
Shapiro, S. L., and S. A. Teukolsky, 1983, *Black Holes, White Dwarfs, and Neutron Stars: The Physics of Compact Objects* (New York: Wiley).
Sher, M. 1989, Physics Reports **179**, 273.
Shu, F. H. 1982, *The Physical Universe* (Mill Valley: University Science Books).
Shu, F. H., F. C. Adams, and S. Lizano, 1987, Ann. Rev. Astron. Astrophys. **25**, 23.
Smoot, G., et al. 1992, Astrophys. J. Lett. **396**, L1.
Stahler, S. W. 1988, Astrophys. J. **332**, 804.
Starobinsky, A. A. 1982, Phys. Lett. 117 B, 175.
Steinhardt, P. J., and M. S. Turner, 1984, Phys. Rev. D **29**, 2162.
Stevenson, D. J. 1991, Ann. Rev. Astron. Astrophys. **29**, 163.
Suzuki, M. 1988, Phys. Rev. D **38**, 1544.
Timmes, F. X. 1996, unpublished.
Tinney, C. G. 1995, Editor, *The Bottom of the Main Sequence and Beyond* (Berlin: Springer).
Tipler, F. J. 1992, Phys. Lett. 286 B, 36.
Tolman, R. C. 1934, *Relativity, Thermodynamics, and Cosmology*, (Oxford: Clarendon Press).
Turner, M. S. 1983, Nature **306**, 161.
Visser, M. 1995, *Lorentzian Wormholes: From Einstein to Hawking* (Woodbury NY: AIP Press).
Voloshin, M. B., I. Yu. Kobzarev, and L. B. Okun, 1975, Sov. J. Nucl. Phys. **20,** 644.

Weinberg, M. D. 1989, *Mon. Not. R. Astron. Soc.* **239**, 549.
Weinberg, S. 1972, *Gravitation and Cosmology* (New York: Wiley).
Weinberg, S. 1977, *The First Three Minutes* (New York: Basic).
Weinberg, S. 1978, Phys. Rev. Lett. **40**, 223.
Weinberg, S. 1980, Phys. Rev. D **22**, 1694.
Weinberg, S. 1989, Rev. Mod. Phys. **61**, 1.
Wilczek, F. 1978, Phys. Rev. Lett. **40**, 279.
Wilczek, F., and A. Zee, 1979, Phys. Lett. 88 B, 311.
Wood, M. A. 1992, Astrophys. J. **386**, 529.
Wright, E. L., et al. 1992, Astrophys. J. Lett. **396**, L3.
Zel'dovich, Ya. B. 1976, Phys. Lett. 59 A, 254.
Zinnecker, H. 1984, Mon. Not. R. Astron. Soc. **210**, 43.
Zuckerman, B., and M. A. Malkan, 1996, *The Origin and Evolution of the Universe* (Sudbury, MA: Jones and Bartlett).

6

Heat Death in Ancient and Modern Thermodynamics

Gábor Kutrovátz

1 The Problem of Heat Death: A Historical Introduction

Aristotle's physics is the most influential physical theory of antique science, and one of the most influential physical theories ever created. The birth of modern science in the seventeenth century is in part a result of the fight against the dogmatic devotion to the Aristotelian scientific mentality, and since then Aristotelian physics has been usually condemned and ridiculed. Before very lately, not many historians of science realized that Aristotle's physics seems to be much more coherent and meritable if we take it out of the context of modern (Newtonian) mechanics, to which it is always compared, and try to see it in its own right [7]. The traditional interpretation is that Aristotle's theory of nature was based mostly on a priori philosophical considerations, and less on empirical observations, therefore it was natural philosophy rather than natural science. Today, however, it is becoming more and more acknowledged that Aristotle's physics is a relevant non-mathematical theory of nature, and that its primary concern is not the set of phenomena that mechanics aims to explain. It seems reasonable to work under the assumption that Aristotle was mostly interested in natural phenomena that are discussed today within the field of thermodynamics [9].

In Aristotle's physics, the problem of Heat Death (as we call it today) is recognized and dealt with, although the way he tackles the question is somewhat alien to our modern mind. For the sake of reconstruction, I will make an analogy between his ideas and the respective theses of modern physics. Aristotle's views can be summarized briefly as the following: *every natural body has its natural place to which it has a tendency to get if no external process prevents it* [11]. (More precisely, this principle concerns the elements of which the bodies consist, and not the bodies themselves, but the above law can be considered as a consequence of the principle, where 'place' is understood in a broader sense than the modern concept of unique spatial location.) This principle raises the following question: Why is there motion in the world, if systems tend toward their 'rest' states? As Aristotle puts it:

> a baffling problem — viz. why the simple bodies, since each of them is travelling towards its own place, have not become dissevered from one another in the infinite lapse of time. [3]

On the other hand, a very similar problem appeared in the nineteenth century with the introduction of the second law of thermodynamics. As the analogy goes, in today's thermodynamics it holds that *every thermodynamic system has its equilibrium state to which it has a tendency to get if no external effect prevents it.* This law was conceptually (although not logically) inconsistent with Newtonian mechanics, the dominant physical theory of that time, since it states that the direction of natural processes is not indifferent. If we extend the validity of the second law to the whole universe as one global thermodynamic system, then we find ourselves in the same situation as Aristotle did: we need to give an explanation to the dynamic, instead of static, state of the universe. Rudolf Clausius introduced the concept of Heat Death with the following words:

> The more the universe approaches this limiting condition in which the entropy is a maximum, the more do the occasions of further changes diminish; and supposing this condition to be at last completely obtained, no further change could evermore take place, and the universe would be in a state of unchanging death. [6]

In this paper, I would like to discuss two very different solutions to the problem given by ancient and modern physics, although I will put a greater emphasis on the modern theory which I discuss in more detail. I hope that this case study can tell us as much about today's cosmological theory as it suggests about the nature of scientific change.

2 The ancient solution

Aristotle's natural world was divided into two regions: *the sublunary region* (consisting of the four material elements of mundane bodies), and *the superlunary region* (consisting of the fifth element of celestial bodies). "Physics", or the "study of nature", was supposed to deal with sublunary phenomena, while the region of celestial spheres was governed by eternal "metaphysical" laws. This fundamental distinction between the two regions of the universe was rigidly sustained in the Aristotelian science of the Middle Ages. According to this picture, our initial problem (of "Heat Death") is relevant only to the natural (sublunary) phenomena. However, Aristotle's final solution did appeal to the eternal celestial world as well, and therefore it can be reconstructed in two stages:

1) The sublunary stage. It is the circular motion of the Sun which constantly induces processes in nature, partly by maintaining the transformation of elements into one another, and therefore by displacing bodies from their natural places:

> For the sun as it approaches or recedes, obviously causes dissipation and condensation and so gives rise to generation and destruction. Now the earth remains but the moisture surrounding it is made to evaporate by the sun's rays

> and the other heat from above, and rises. Vapour cools because its heat is gone and because the place is cold, and condenses again and turns from air into water. And after the water has formed it falls down again to the earth. [. . .] So we get a circular process that follows the course of the sun. [2]

It is worth mentioning that this quote eloquently illustrates why most of Aristotle's physics can be interpreted as some sort of proto-thermodynamics, rather than proto-mechanics. His explanations very often appeal to the four mundane elements (earth, water, air, fire), and it leads to a great number of discussions about heat phenomena.

2) The superlunary stage. In spite of the "irreversibility" of Aristotle's natural processes, the eternal motion of natural bodies (contrary to their 'nature') is maintained by the eternal motion of celestial spheres (inherent in them), namely that of the Sun's sphere. For as the Sun is moving around the Earth, it constantly changes the heat conditions of natural processes. Thus, postulating this inherent eternal motion of the spheres would suffice for the purpose of answering the original question concerning motion of natural bodies. But Aristotle continues his chain of causation. The motion of the Sun's sphere is brought about by the motion of outer spheres (each of them causing the inner neighbour to move), and the motion of the outmost sphere is caused by the First (or Unmoved) Movent.

> [. . .] for it is impossible that there should be an infinite series of movents, each of which is itself moved by something else, since in an infinite series there is no first term — if then everything that is in motion is moved by something, and the first movent is moved but not by anything else, it must be moved by itself. [4]

This is a purely a priori argument concerning the chain of causation: if all bodies are moved by other bodies then, at the beginning of this chain, there must be an actor that moves but is not moved. (As many ancient Greek thinkers, Aristotle was against, and could not accept, the notion of infinity.) That this First Movent must be located at the outmost heavenly sphere, is 'proved' in the following way:

> Moreover the [first] movent must occupy either the centre or the circumference, since these are the first principles from which a sphere is derived. But the things nearest the movent are those whose motion is quickest, and in this case it is the motion of the circumference that is the quickest: therefore the movent occupies the circumference [of the world]. [5]

This is in part another a priori argument, although it appeals to the observation that bodies which are directly caused to move are quicker in their motion than bodies caused to move indirectly (or in a mediated way). Aristotle's physics is based, on the one hand, on the observation of everyday phenomena (where friction and dissipation are always present, unlike in Newton's 'ideal cases' where conservation laws can be formulated) and, on the other hand, on more general philosophical principles about nature and being. Physics and metaphysics are not separated in ancient theories of nature; but are they separated in modern physics?

3 The modern solution

The nineteenth century problem of Heat Death was taken its sting out in the twentieth century, with the birth of modern cosmology. Since the expanding universe is not an eternal entity, there is no direct contradiction between the universal validity of the second law of thermodynamics and the empirical evidence of natural processes. Even if the final state of Heat Death is unavoidable, since the past history of the universe is finite, there is nothing paradoxical in that we have not reached this state yet: it will happen some time in the future. The so-called "Heat Death paradox" is thus resolved.

However, this solution raises another problem: if the preferred state of systems is equilibrium, why are they not in their equilibrium state today? What happened in the past course of cosmic development which can be called for explaining this non-equilibrium picture of the world? In other words, our initial question is still valid: Why is there motion in nature? Or more precisely: Where does the existing motion originate from?

3.1 Extropy

In order to formulate the question in a language which is more suitable for modern scientific purposes, let us introduce the notion of *extropy* [10]. It is a state parameter which expresses the distance of a system from its equilibrium state on the entropic scale. In other words, it measures the entropy production of the imaginary process that takes the system to equilibrium with the environment. It is the difference between the maximal entropy of system plus environment and the actual entropy of system plus environment,

$$\Pi = S_e^\tau + S_e^k - S^\tau - S^k. \tag{1}$$

In the case of the universe, the "environment" does not exist (since the universe contains all entities that can be expressed in the physical language), so the extropy of the universe is expressed in a slightly different form,

$$\Pi^U = S_e^U - S^U. \tag{2}$$

A system is in equilibrium if its extropy is zero: in the case of the universe, this is what we call 'Heat Death'. Our question, therefore can, be formulated in the following way: What process can produce global extropy in the cosmic scale?

3.2 The model universe

Let us regard the universe as a system which is determined by three extensive state parameters [1]:

1. For *R*:

$$\frac{1}{2}\dot{R}^2 = \frac{G}{R}\left(\frac{4\pi}{3}R^3\rho\right) + \kappa, \tag{3}$$

where *R* is the 'radius' of the universe (the scale parameter), *G* is the gravitational constant, ρ is the matter density, and κ is a constant that determines whether the universe is open ($\kappa \geq 0$) or closed ($\kappa < 0$).

2. For *E*:

$$\dot{E} + p\dot{V} = 0, \tag{4}$$

where *E* is the internal energy of the universe, *p* is the pressure, and *V* is the volume (proportional to R^3). We can see here that the internal energy is not constant because it is transformed into the energy of the gravitational field: this is a peculiar feature of the universe as a thermodynamic system.

3. We have a third extensive parameter in the description: *N* is the number of particles in the universe. Unlike in the case of the other two extensive parameters, we suppose that *N* is constant.

In this model universe we assume the presence of two matter components: atomic hydrogen and black-body photon gas, both homogenous and isotropic. We take that these components must cool as the universe expands. However, the cooling rates of the two components obey different laws: the baryonic matter would cool according to R^{-2}, while the photon gas would cool according to R^{-1}, if the only thermodynamic process present was the expansion of the universe. In other words, the temperatures of the components tend to be different, and there is a process of heat flow from the hotter to the cooler component. In our model universe, the strength of this interaction is given as an arbitrary numerical value (considered to be constant).

We can now derive a system of differential equations (given in [1]) which, assuming that the radius changes according to Eq. 3 above, enables us to determine the behaviour of the following state parameters of the system: *R,* the radius of the universe; T_g, the temperature of the gas component; T_r, the temperature of the radiation component; and *S,* the total entropy of the universe. The system of equations is the following:

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{2c}{R^2} + 2c_1\gamma T_r^4 + \gamma B R^{-3}\left(2m + 3T_g\right), \tag{5}$$

$$\frac{\dot{T}_r}{T_r} + \frac{\dot{R}}{R} = \frac{1}{4c_1} A B T_r^{-1} R^{-6}\left(T_g - T_r\right) \tag{6}$$

$$\frac{\dot{T}_g}{T_g} + 2\frac{\dot{R}}{R} = \frac{2}{3} A T_r^3 T_g^{-1}\left(T_r - T_g\right) \tag{7}$$

$$\dot{S} = ABR^{-3}T_r^2T_g^{-1}\left(T_r - T_g\right)^2, \tag{8}$$

where B is the baryonic number, m is the molecular mass, A is the coefficient of the interaction, C_g and C_r are the heat capacities of the gas and radiation components, respectively, c and c_1 are constants, and the rest has been explained above.

Given the values of the parameters yielded by the solution of the system of equations, we can easily calculate the extropy function of the model universe. That is, we calculate the amount of entropy the system would produce by getting to equilibrium at a given instant of time. Provided that the heat capacities are constant, the expression of the extropy is

$$\Pi = \left(C_r + C_g\right)\ln\frac{T_g + DT_r}{\left(1+D\right)T_g^{\frac{1}{1+D}}T_r^{\frac{D}{1+D}}} \tag{9}$$

where $D = C_r/C_g$.

The development of the model universe was simulated numerically by the program *Mathematica*.

3.3 Evolution of the model universe

We have to distinguish between two qualitatively different cases: If the universe is open ($\kappa \geq 0$) then the expansion will continue for ever (because the gravity of the massive matter is not strong enough to stop it) and the universe has an infinite future; but if it is closed ($\kappa < 0$) then the expansion will stop and turn to contraction, by the end of which the history of the universe is finished. The evolution paths of state parameters belonging to an open universe are plotted in Figure 6.1, while the case of a closed universe is shown in Figure 6.2.

On the left side of the figures one can see the graphs of the radius, the temperatures of components, and the entropy. On the right side one finds entropy, extropy, and the maximum value of entropy ($S_{max} = \Pi + S$) belonging to the equilibrium state at the given extensive parameters. The strength of the interaction between the components (A) increases from the top figures towards the bottom ones, the value being 0 in the first case and very large in the bottom case.

Let us concentrate on the case of open universe first. We see that two factors put a constrain on the value of entropy: if the interaction is weak, there is obviously not much entropy produced; while if the interaction is strong, the temperature difference is small, so the interaction process becomes weak again. There is an optimal case for entropy production when the universe is the most 'active' thermodynamically. Nonetheless, extropy behaves in a different way: the dependence of extropy on the strength of interaction follows a monotonous function. The stronger the interaction the smaller the extropy becomes, and the closer the system gets to equilibrium.

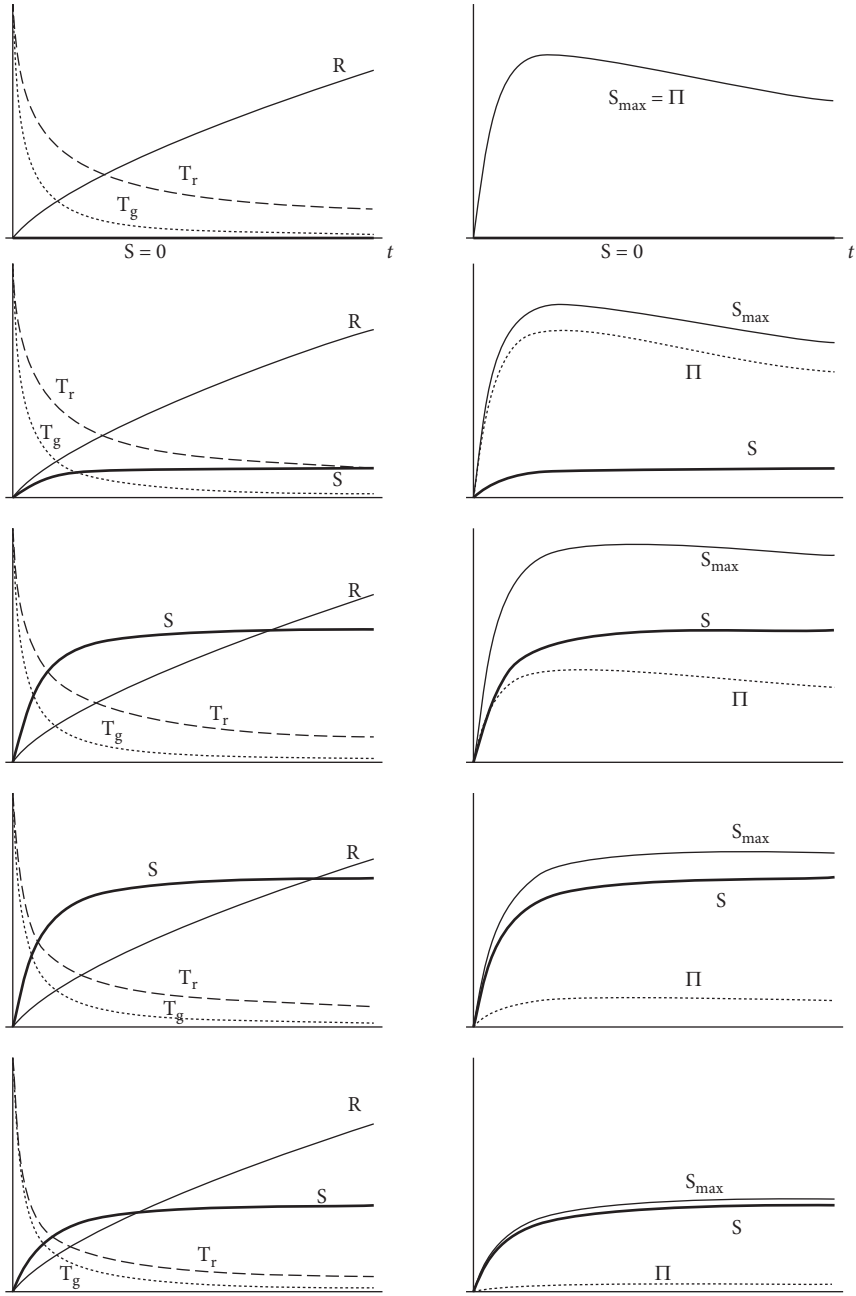

Figure 6.1 Thermodynamic functions in an open universe.

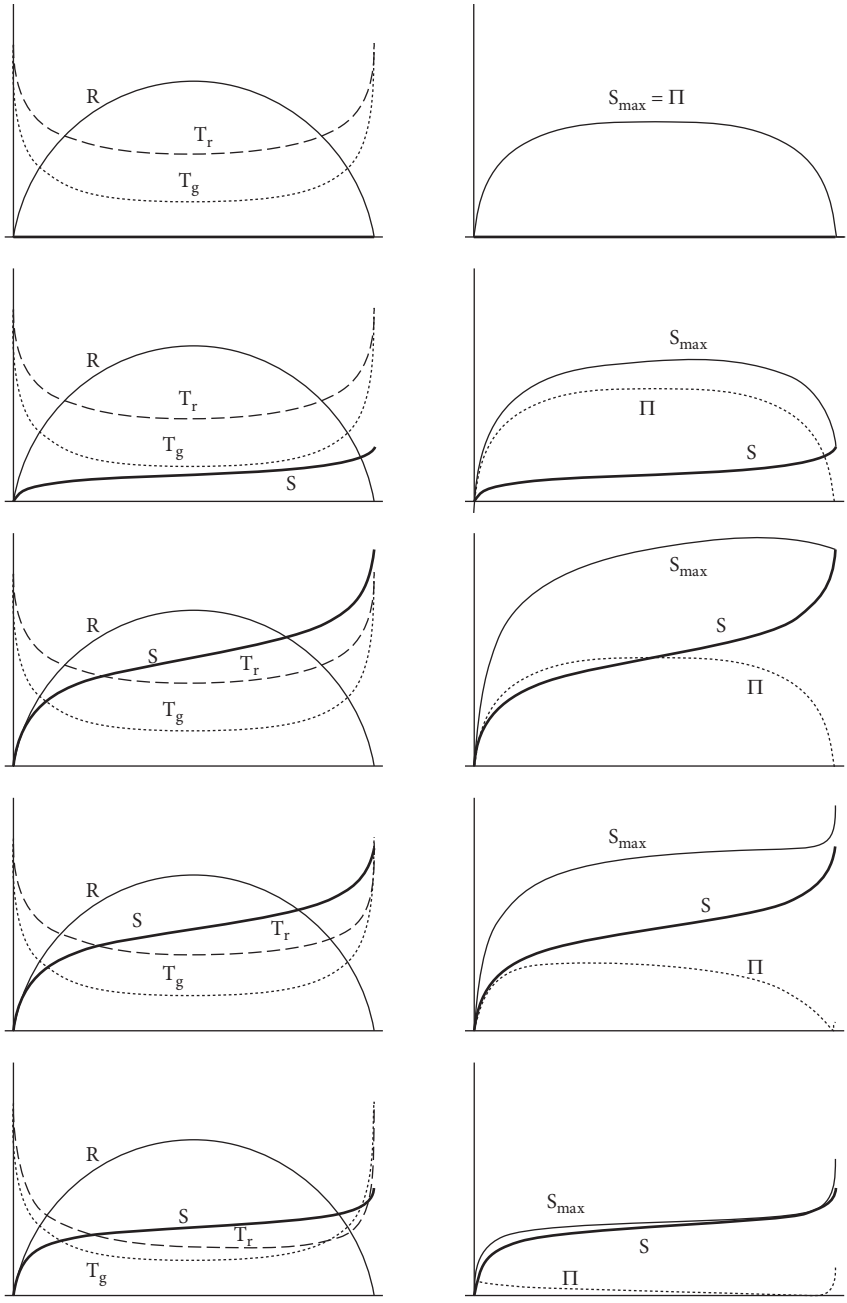

Figure 6.2 Thermodynamic functions in a closed universe.

Let us now turn to the time-dependence of the parameters. We can see that no matter how strong the interaction is, the extropy function follows the same qualitative route with time: it increases in the beginning (until the temperature difference becomes maximal), but then it starts to decrease and tends to zero as time approaches infinity. It means that, in our open model universe, there is no way of avoiding the final state of Heat Death, which is the boring state of absolute, eternal smoothness and peace. Still, thanks to the expansion of the universe, the temperatures of the two components become different vehemently in the beginning, at the very early stages of cosmic history. Therefore, there is an initial increase in the value of extropy as it becomes greater than zero and the universe leaves its equilibrium state. Thus the expansion of the universe creates the possibility of all the physical processes of the future. To put it very simply, extropy means process in physical systems.

In the case of a closed universe, Heat Death becomes a marginal problem: since the future of the universe is finite, the equilibrium state may or may not occur before the end. But it is important to notice that equilibrium in a contracting universe is *not* an endless state of Heat Death, but it is a momentary phase. As the components, owing to the contraction, start to warm up again the temperatures get closer to one another. But if the interaction is not negligible then the contraction is not symmetrical to the expansion, because the direction of the heat flow does not turn back when the collapse begins. As a consequence, the plots of the two temperatures cross each other before the end, and this results in an immense final extropy increase. Just before the very end of everything, the universe becomes 'active' again for a short while.

3.4 Beyond the model

Of course, if we want to give a more accurate answer to the question whether Heat Death will occur or not, we have to transcend the limits of our simple model. As we saw, the problem of Heat Death is relevant only in an open universe, so in this section I will constrain myself to this case and suppose that the universe has an infinite future.

Let us first turn to the assumption that the components are homogeneously distributed. While the $2.7K$ background radiation (the photon gas) is proved to be practically homogeneous indeed, baryonic matter is condensed into different material structures discontinuous from one another: stars, galaxies, clusters of galaxies, etc. Condensation increases internal energy in these dense regions of matter, and thus baryonic structures significantly hotter than the photon gas are born. However, these structures radiate their heat surplus into the photon gas (bringing about small inhomogeneities in it), and even if this radiation process is enormously lengthened by nuclear reactions inside stars as sources of heat, this temperature reversion between regions of the two components is only a temporary phenomenon. All stars — in fact all condensations of baryonic matter – must use up their energy supply sooner or later. Even black holes, first thought to violate the second law of thermodynamics, evaporate through quantum gravity effects and therefore have a finite, although sometimes very long, life span before they disappear. From the viewpoint of Heat Death, the present inhomogeneous picture of the universe, as we know it today, is but an incidental contingency.

Another oversimplification of our model is that it considers only two components. But there seems to be only one significantly massive component in addition to the two already discussed: the neutrino gas. This follows the same cooling rate as the photon gas does, although they are not in equilibrium (the photon gas was heated by the $e^+ - e^-$ annihilation at the very early stages). The neutrino gas, however, does not interact with the photon gas, and its interaction with the baryonic matter is also very weak, so its expansion is practically adiabatic: it plays hardly any role in the overall thermodynamic picture of the universe. Furthermore, the qualitative results derived from the model do not change even if there are more components present: all the temperatures are ever decreased by the expansion of the universe. Concerning the future of the two known components, protons, as we know it today, have a finite life span and after their decay the free neutrons will disappear almost immediately. In other words, baryonic matter will vanish from the universe. Whether neutrinos will remain forever or they will decay too is not determined yet but, as we have seen, they do not matter much from the viewpoint of Heat Death.

Having mentioned the early stages of the universe, let us examine how phase transitions of matter can influence our results. When the temperature of the universe drops below a value $T_x = m_x c^2/k$, where m_x is the mass of a certain type of particle, then this particle enters into its non-relativistic phase and realises a new material component in the universe (e.g. quarks-antiquarks, neutrinos-antineutrinos, electrons-positrons). Now, this phase transition clearly increases the global entropy of the universe [8], but how about extropy? The total entropy available for the whole system also increases (following simple considerations in statistical physics), and this increase is never smaller (and generally larger, except when the components contain equal numbers of particles) than the rise in actual entropy. That is, extropy grows in phase transitions: the universe gets further from thermal equilibrium.

Now, it seems today that the era when phase transitions occurred is long gone: this was the hot and dense state when events followed one another rapidly. However, we do not know for sure if such drastic changes will not happen in the future, because particles with very small mass might appear at still lower temperatures. Still, we have a very good reason to suppose that there occurs (or occurred) once a last transition: since the universe is finite in mass/energy, there can be only a finite number of material components in it. Phase transitions cannot postpone Heat Death for ever.

Finally we can see that the whole problem boils down to the following simple conceptual scheme applied in these cases: since the universe is a finite system there can be no infinite process whatsoever (known or unknown) that can keep it out of equilibrium state forever. But this is, as we can call it, a 'metaphysical' principle and, except for the finiteness of the universe, no empirical knowledge is embedded in it. In order to explain, through particular physical processes, why the universe will not escape final equilibrium we should examine from case to case an infinite number of conceivable possibilities and rule them out one by one, which is impossible. Again, metaphysical (or, more or less, normative) principles seem to 'intrude' unavoidably to our scientific knowledge.

4 Conclusion

We have seen two very different solutions to the problem why the observed 'activity' of nature do not contradict the irreversibility of physical processes. In the ancient case, Aristotle sacrificed the closedness of the universe and put a theoretical entity at the border of the world, which plays as an eternal source of motion. Today cosmological theory sacrificed the belief in an endless, eternal universe and postponed Heat Death into the indeterminate future, regarding the present picture of the world as contingent and temporary. However, modern theory cannot formulate certain claims about Heat Death: this question must be re-examined in the light of all newly discovered effects and processes. As a concept concerning eternity, Heath Death seems to belong, not so much to the field of our scientific knowledge, but to the horizon and limits of our prevailing theories about the universe.

Bibliography

[1] R. A. Alpher, Gy. Marx, Vistas in Astronomy **35**, (1992).
[2] Aristotle, *Meteorologica*, 336b-347a, Clarendon Press, Oxford, 1907.
[3] Aristotle, *De generatione et corruptione*, 289a, Clarendon Press, Oxford, 1953.
[4] Aristotle, *Physics*, 256a, Harvard University Press, Oxford, 1957.
[5] Aristotle, *Physics*, 269a, Harvard University Press, Oxford, 1957.
[6] R. Clausius, Phil. Mag. **4**, 405 (1868).
[7] Th. S. Kuhn, *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago, 1962.
[8] P. T. Landsberg, From entropy to God?, in: *Thermodynamics: History and Philosophy*, K. Martinás, L. Ropolyi, and P. Szegedi, eds., World Scientific, Singapore, 1990.
[9] K. Martinás, Aristotelian Thermodynamics, in: *Thermodynamics: History and Philosophy*, K. Martinás, L. Ropolyi, and P. Szegedi, eds., World Scientific, Singapore, 1990.
[10] K. Martinás, Per. Polytech. S. Chem. Eng. **42**, (1998).
[11] K. Martinás, Irreversibility in Aristotelian Physics, in: *Volume of Abstracts: 10th LMPS Congress*, Florence, 1995.

7

Entropy and Eschatology: A Comment on Kutrovátz's Paper "Heat Death in Ancient and Modern Thermodynamics"

Milan M. Ćirković

1 Introduction

The problem of heat death and the general issue of the thermodynamical temporal asymmetry is one of the long-standing and ever-inspiring topics in the history of human philosophical and scientific thought, comparable to Zeno's paradoxes on space and motion or the Hicetas-Berkeley-Mach puzzle of the origin of inertia of matter. Recent discussion of the topic by Kutrovátz [22] in this Journal is indicative of the great surge of interest in physical, philosophical as well as the historical aspects of this and related themes (for some other recent discourses, see [3, 24, 28, 29, 37, 38]. It brings (i) a very welcome appraisal of the state of the problem in by far the most influential ancient physical theory—Aristotelian physics—and (ii) unearths some of the valuable comparison between that state of affairs and the situation in modern cosmological thought. However, its treatment is incomplete in several important points, and contains—especially in the latter part-some (widely spread) misconceptions and mispresentations. Our goal in this note is, therefore, to correct some omissions in Kutrovátz's manuscript, as well as to suggest generalizations of its conclusion in the light of the nascent astrophysical discipline of physical eschatology. This seems an excellent case study for emphasizing the degree our cosmological knowledge is determined by the underlying information flow. In addition, we point out another, surprisingly modern, answer to the heat death puzzle which existed in the antiquity, in a tradition different from the Aristotelian one. We begin with the latter topic.

2 Alternative answer to the heat death problem

In contradistinction to the Aristotelian "prime movent" which is eternally copresent with the world, the modern evolutionary idea that the universe is of finite age and

is currently unwinding as a consequence of finite amount of primordial motion was also present in antiquity, even prominent in some circles. Unfortunately, those were the circles of philosophers of nature who were either agnostics or open atheists, and therefore were carefully suppressed and neglected in subsequent centuries. The elements of such a worldview were present in cosmologies of Anaximandros, Empedocles, Anaxagoras and Democritus, as well as in some later epigones. We consider here briefly some of these ideas.

In the very first chapter of the immortal history of Thucydides, there is a famous statement that before his time—i.e. about 450 BC—nothing of importance (*ου μεγαλα γενεσθαι*) had happened in history. This startling statement has been correctly called "outrageous" by Oswald Spengler, and used to demonstrate the essentially mythological character of the ancient Greek historiography [32, 7]. It may indeed be outrageous from the modern perspective, but it does motivate a set of deeper questions, ultimately dealing with cosmology. The fact that Thucydides did not know (or did not care to know) previous historical events, does not change the essential perception of *finiteness* of human history inseparable from the Greek thought. This property starkly conflicts with the notion of an *eternal continuously existent world,* as it was presented in both modern and ancient cultures. Obviously, it is irrelevant which exact starting point we choose for unfolding historical events. In any case, the number of events is finite, and the time span considered small, even compared to the specific astronomical timescales (some of which, like the precession period of equinoxes, were known in the classical antiquity, as is clear from the discussion in *Timaeus*), not to mention anything about a past temporal infinity. Although there was no scientific archaeology in the ancient world, it was as natural then as it is now to expect hypothetical previous civilizations inhabiting Oikumene to leave some traces—in fact, an infinite number of traces for an eternally existent Oikumene! There are indications that pre-Socratic thinkers have been aware of incompatibility of this "Thucydidean" finiteness of historical past with the eternal nature of the world. We have already mentioned the solution (periodic singular states) proposed by Empedocles himself. Even earlier, in the fragmentary accounts of the cosmology of Anaximandros [i.e., Anaximander], one may note that he proposed an evolutionary origin of humankind in some finite moment in the past, parallel to his basic postulate of separation of different worlds from apeiron and their subsequent returning to it.[1] In Anaxagoras' worldview, there is a famous tension between the eternity of the world's constituents and the finite duration of movement (and, therefore, relational time) in the world. In the same time, it seems certain that Anaxagoras, together with Anaximandros and Empedocles, was an early proponent of the evolutionary view, at least regarding the origin of humankind [16].

Finally, an almost modern formulation of the anthropic argument for the finite past has been made in Roman times by Lucretius, who in Book V of his famous poem *De Rerum Natura* wrote the following intriguing verses:

> Besides all this,
> If there had been no origin-in-birth

Of lands and sky, and they had ever been
The everlasting, why, ere Theban war
And obsequies of Troy, have other bards
Not also chanted other high affairs?
Whither have sunk so oft so many deeds
Of heroes? Why do those deeds live no more,
Ingrafted in eternal monuments
Of glory? Verily, I guess, because
The Sun is new, and of a recent date
The nature of our universe, and had
Not long ago its own exordium.[2]

For highly scientific-minded Lucretius, the shortness of human history *is* very strange on the face of hypothesis of the eternal existence of the world. Although the references to "eternal monuments" and "other bards" sound naive, it is clear that he had in mind any form of transmission of information from the past to the present; and an infinite amount of information from an infinite past. His empirical assessment of the surrounding world clearly shows the absence of such information. Therefore, an explanation is needed. The simplest explanation, as Lucretius was highly aware, is to treat the argument as *reductio ad absurdum* of the starting hypothesis (eternal nature of the world) and to assume that the world is of finite—and relatively small—age.

The depth of Lucretius' thought in this passage is almost amazing, especially when the historical blindness of subsequent generations to this same argumentation is taken into account. Lucretius' argument applies to the classical Newtonian universe of infinite age, as well as to modern stationary alternatives to the evolutionary cosmology (like the classical steady-state theory). It emphasizes the technological nature of possible evidence ("ingrafted in monuments"). This is exactly what modern cosmologists Davies and Tipler have had in mind when constructing the anthropic argument in order to refute the eternal cosmologies of our epoch. Lucretius' monuments play essentially the same role as Tipler's von Neumann probes sent by advanced intelligent communities [34]. Their absence testifies on the finite past. We conclude that a rather modern idea of the (relatively) recent origin of the universe and its "running down" has been present in classical antiquity. Of course, it did never attain the attention and considerations accorded to the Aristotelian picture, but it has been present nevertheless, testifying upon the high degree of scientific sophistication of the ancient world. It has been resurrected in its modern form, for instance, in the famous debate on the thermodynamical recurrence between Boltzmann and Zermelo (cf. [33]), which is the first instance scientific (in the modern sense) cosmological speculations appeared in a respected peer—reviewed research journal. In the course of [the twentieth] century, it has been incorporated into the standard cosmological lore, in particular after the victory of evolutionary models over their steady-state rival in the "great battle" of 1950s and 1960s [19]. It is exactly this lore we now turn to.

3 Modern cosmology and some errors in Kutrovátz's account

In this section we list and briefly consider several important astrophysical points missing in Kutrovátz's account of the modern answer to the heat death puzzle. These considerations will certainly help in highlighting the entire scope of the problem whose important aspects, we wholeheartedly agree with Kutrovátz, are still open.

3.1 Gravitational field as the major source of entropy

The basic missing part in Kutrovátz's toy model of the entropy evolution of the universe is gravitational entropy. It has very slowly dawned of physicists and cosmologists that, apart from the thermodynamical entropy, gravitational field may store a huge quantity of the internal degrees of freedom, thus having potentially huge entropy in itself. This has been realized in the black hole context in early 1970s, with the revolutionary studies of Bekenstein [5] and Hawking [17]. Bekenstein-Hawking formula gives us a hold on the entropy of gravitational field of a black hole, and there certainly are ways of discerning gravitational entropy in other cases, although they are still highly speculative (being dependent, of course, on the structure of correct quantum theory of gravitation!). Introducing gravitational entropy enables solving several important puzzles in astrophysics and cosmology (e.g. [26]). The most important of them is the state of thermodynamical equilibrium of very early cosmological epochs, as discerned from the high isotropy of the cosmological black-body radiation. If the classical thermodynamical entropy is all there is, how then could anything occur in the universe after the epoch of recombination at redshift $z \sim 1500$? We would naively expect plasma to simply reach detailed equilibrium and remain in such state forever. Instead, gravitational clumping acted to reduce the thermodynamical entropy at the expense of the gravitational entropy which greatly increased during the process of cosmological structure formation. And cosmological structure formation, in turn, enabled all the wealth of subsequent physical, chemical and biological processes we are now dealing with in the sciences; it is, among other things, the cause of our appearance on Earth as intelligent observers.

3.2 Thermodynamical capacities of gravitating systems

When Kutrovátz writes that "we can see here that the internal energy is not constant because it is transformed into the energy of the gravitational field: this is a peculiar feature of the universe as a thermodynamic system", he is only partially correct. While this undoubtedly holds true for the universe, it is also true for many other astrophysical systems, notably stars and galaxies. In fact, it is true for any system that is held together by a long-range force like gravity. Suppose that we add energy to a star (say, by a giant laser beam), and then wait some time for the relaxation processes to occur. We shall notice that its size has increased and its temperature will actually decrease (familiar

example of the negative heat capacity). This may seem irrelevant for the particular cosmological case. However, the conclusion often drawn from the same property when applied to the Hubble expansion, namely that the Hubble expansion is the cause of the entropy gradient (i.e. that expansion creates new extropy), is wrong. This point has been the focus of a fierce controversy raging in 1960s and 70s, beginning with the classical study of Thomas Gold ([15]; see also [8, 23]). The idea was to explain the departure from thermodynamical equilibrium assuming that the universal expansion creates new states for new configurations of matter, so that the entropy of matter begins to lag more and more behind the maximal possible entropy. Ingenious as it was, this idea has been abandoned since for several reasons, the main being that the "special" low entropy nature of the initial Big Bang singularity is so exceptional (as calculable in principle from Boltzmann formula $S = k \ln W$; see [26]) that the amount of subsequent expansion produces almost negligible effect. In other words, we still need to explain very low initial *gravitational* entropy, whose subsequent increase more than offsets the apparent (thermodynamical) entropy decrease after the recombination epoch.

3.3 Open vs. closed universes and cosmological constant

In several places in Kutrovátz's paper [22] the confusion of open/closed vs. ever-expanding/recollapsing is perpetuated. This has become very widespread as a consequence of the general neglect (in particular in textbooks) of models with cosmological constant (or any other form of "dark energy") in the approximate 1930–1990 period. Today, after spectacular results of the recent cosmological observations of distant supernovae, as well as of anisotropies of the microwave background radiation, we are in much better position. For a nicely written summary of the situation at present see [20].

Let us summarize some cosmological basics. Ω is the ratio of physical density of all matter fields to the so-called critical density necessary for universe to stop current expansion and recollapse toward Big Crunch. In the misleading textbook discourse which Kutrovátz uncritically accepts, $\Omega \leq 1$ universes will expand forever, while $\Omega > 1$ universes will recollapse. This strictly applies only to the case of matter fields possessing "regular" equation of state; in the presence of vacuum energy, indicated by recent cosmological supernovae experiments, the situation becomes more complicated. There, we can write $\Omega = \Omega_m + \Omega_\Lambda$, $\Omega_\Lambda$being the contribution of dark energy. Now, even $\Omega > 1$ (= topologically closed) universes may expand forever, under the condition that the sign of dark energy is positive (corresponding to the repulsive effective force). This will occur for any $\Omega_m \leq 1$, and if $\Omega_m > 1$ for (e.g. [6])

$$\Omega_\Lambda \geq 4\Omega_m \cos^3\left[\frac{1}{3}\arccos\left(\frac{1-\Omega_m}{\Omega_m}\right)+\frac{4\pi}{3}\right]$$

Thus we have a degeneracy in relationship of topological properties (open vs. closed) and its dynamical future (ever-expanding vs. recollapsing). With positive cosmological constant (which seems realistic; e.g. [27, 31, 36]) topologically closed universe may also

expand forever, especially if its contribution is so large as indicated by observations ($\Omega_\Lambda \approx 0.7$). We have no observational way of determining whether the real universe is open or closed, since — and especially if one believes in inflationary models — the total cosmological density is $\Omega = 1 \pm \varepsilon$, ε being of the order of 10^{-5}. However, we may be virtually certain that the universe, no matter its topological properties, will expand forever.

3.4 Extropy in open universes and the definition of the heat death

If the definition of heat death is — as classically always assumed — the state of reached maximal entropy of a closed physical system, than the issue whether heat death will occur in an ever-expanding universe is not trivial or obvious. Although entropy may increase forever, this does not mean it will reach maximal value, if that maximal value can recede indefinitely. In fact, it seems that for the classical Einstein-de Sitter universe (matter-dominated universe with $\Omega = \Omega_m = 1$) exactly that is the case: in one of the first studies of cosmological future (appropriately entitled "Entropy in an Expanding Universe"), Frautschi [14] wrote:

> It is apparent . . . that the entropy in a causal region falls steadily further behind max S during most of the cosmic history. S/S_{max} does increase temporarily during the period of stellar and galactic black hole formation. Life as we know it develops during the same period, utilizing the much smaller, but conveniently arranged entropy generation on a planet or planets situated near nucleosynthesizing stars. . . . the expanding universe does "die" in the sense that the entropy in a comoving volume asymptotically approaches a constant limit.

This is *not* the heat death as classically understood (e.g. [12]). In the similar vein, Adams and Laughlin [1] in the comprehensive study of almost all aspects of physical eschatology notice that

> Thus far in this paper, we have shown that entropy can be generated (and hence work can be done) up to cosmological decades $\eta \sim 100$. [Cosmological decades are defined as epochs of time $t = 10_\eta$ years] . . . The formation of larger and larger black holes, can continue as long as the universe remains spatially flat and the density perturbations that enter the horizon are not overly large. The inflationary universe scenario provides a mechanism to achieve this state of affairs, at least up to some future epoch . . . Thus the nature of the universe in the far future $\eta \gg 100$ may be determined by the physics of the early universe (in particular, inflation) at the cosmological decade $\eta \sim -45$.

But it is not necessary to enter into wealth of technical details of these and related studies here; it is enough to point out that there already exists a sizeable volume of scientific literature on the entropy production in the cosmological future — literature Kutrovátz unfortunately ignores.

On the other hand, in the realistic models with the positive cosmological constant, heat death may be operationally defined as the state of reached maximal entropy *within the event horizon.* In this case the assertion of Kutrovátz that the heat death is inescapable is correct.

4 Discussion: Physical eschatology and open systems

Physical eschatology is a rather young branch of astrophysics, dealing with the future fate of astrophysical objects, as well as the universe itself [1, 11, 30, 35]. Some relevant issues have been discussed in the monograph of Barrow and Tipler [4], as well as several popular-level books [2, 9, 18]. Since the distinction between knowledge in classical cosmology and physical eschatology depends on the distinction between past and future, several issues in the physics and philosophy of time are relevant to the assessment of eschatological results and *vice versa.*

At a first glance, it seems that — in contradistinction to most other physical disciplines — physical eschatology is at least fortunate enough to deal with the exemplary closed system: the universe itself. But there is one crucial sense in which physical eschatology is — contrary to this impression — a study of open systems. This is nicely manifested in discussion of the heat death of the universe. A necessary ingredient in most serious discussions of physical eschatology is presence of living and intelligent systems in future of the universe (which *ex hypothesi* did not exist in its past). Dyson has been the first to boldly spell it out in 1979:

> It is impossible to calculate in detail the long-range future of the universe without including the effects of life and intelligence. It is impossible to calculate the capabilities of life and intelligence without touching, at least peripherally, philosophical questions. If we are to examine how intelligent life may be able to guide the physical development of the universe for its own purposes, we cannot altogether avoid considering what the values and purposes of intelligent life may be. But as soon as we mention the words value and purpose, we run into one of the most firmly entrenched taboos of twentieth-century science.

The future of universes containing life and intelligence is *essentially* different from the future of universes devoid of such forms of complex organization of matter, as well as different from the past of the same universes in which complexity was lower. In a similar vein, John A. Wheeler wrote in a beautiful paper on the relationship of quantum mechanics and cosmology [39]:

> Minuscule though the part is today that such acts of observer-participancy play in the scheme of things, there are billions of years to come. There are billions upon billions of living places yet to be inhabited. The coming explosion of life opens the door to an all-encompassing role for observer-participancy: to build, in time to come, no minor part of what we call *its* past—*our* past, present and future—but this whole vast world.

Taking into account intentional actions of intelligent beings transforms this field into a study of systems open to the interaction with such advanced noospheres (particularly in light of our virtually complete ignorance of the physics of conscience and intelligence at present). The pioneering study of Dyson quoted above presents an excellent example (although it may be wrong in quantitative details; see [21]) of such a research. In it Dyson envisages a way for advanced intelligent communities to avoid heat death via ingenious methods of energy conservation and information flow optimization. Dyson's work provoked a lot of subsequent research activity. Although we cannot enter into this discussion here, the general moral that future of the universe, including the possible heat death, has recently entered the physical—instead of metaphysical—domain seems inescapable. This stands in a stark contrast to the course of thinking about the cosmological future in previous epochs which Kutrovátz describes, and heralds a completely new era in our understanding of the universe.

Acknowledgments

The author wishes to acknowledge Prof. Branislav Nikolić and Vesna Milošević-Zdjelar for their kind help in finding several key references. Special thanks are owed to Nataša Bulut for her invaluable inspiration and wholehearted support.

Notes

1 This is clear, for instance, from the fragment A 10 in [10], preserved by Plutarch, in which it is explicitly asserted that formation and destruction of many worlds occurs within the global temporal infinity. In the continuation of the very same excerpt from Stromateis, an evolutionary doctrine is attributed to Anaximandros: ". . . Farther he says that at the beginning man was generated from all sorts of animals, since all the rest can quickly get food for themselves, but man alone requires careful feeding for a long time; such a being at the beginning could not have preserved his existence" [13]. Hyppolites quotes Anaximandros as emphasizing the nature of apeiron as eternal (B 2), obviously in opposition to mankind, which has a fixed beginning in time. Even more intriguing is the doctrine ascribed to Anaximandros by Cicero: "It was the opinion of Anaximandros that gods have a beginning, at long intervals rising and setting, and that they are the innumerable worlds. But who of us can think of god except as immortal?" Did he have in mind essentially what we today in SETI-related discussions denote as supercivilizations (e.g. [4])?

2 In translation of William E. Leonard, available via WWW Project Gutenberg [25].[2]

Bibliography

[1] F. C. Adams and G. Laughlin, Reviews of Modern Physics **69**, 337 (1997).

[2] F. C. Adams and G. Laughlin, *The Five Ages of the Universe*, The Free Press, New York, 1999.

[3] D. Albert, *Time and Chance*, Harvard University Press, Cambridge, 2000.
[4] J. D. Barrow and F. J. Tipler, *The Anthropic Cosmological Principle*, Oxford University Press, New York, 1986.
[5] J. D. Bekenstein, Physical Review D **7**, 2333 (1973).
[6] S. M. Carroll, W. H. Press, and E. L. Turner, Annual Review of Astronomy and Astrophysics **30**, 499 (1992).
[7] F. Cornford, *Thucydides Mythistoricus*, Greenwood Press Publishers, New York, 1965.
[8] P. C. W. Davies, *The Physics of Time Asymmetry*, University of California Press, Berkeley, 1974.
[9] P. C. W. Davies, *The Last Three Minutes*, Basic Books, New York, 1994.
[10] H. Diels, *Presocratic Fragments*, Naprijed, Zagreb, 1983.
[11] F. J. Dyson, Reviews of Modern Physics **51**, 447 (1979).
[12] A. S. Eddington, Nature **127**, 447 (1931).
[13] A. Fairbanks, *The First Philosophers of Greece*, K. Paul, Trench & Trubner, London, 1898.
[14] S. Frautschi, Science **217**, 593 (1982).
[15] T. Gold, American Journal of Physics **30**, 403 (1962).
[16] W. K. C. Guthrie, *A History of Greek Philosophy II*, Cambridge University Press, London, 1969.
[17] S. W. Hawking, Nature **248**, 30 (1974).
[18] J. N. Islam, *The Ultimate Fate of the Universe*, Cambridge University Press, Cambridge, 1983.
[19] H. Kragh, *Cosmology and Controversy*, Princeton University Press, Princeton, 1996.
[20] L. M. Krauss and Turner, General Relativity and Gravitation **31**, 1453 (1999).
[21] L. M. Krauss and G. D. Starkman, Astrophysical Journal **531**, 22 (2000).
[22] G. Kutrovátz, Open Systems and Information Dynamics **8,** 349 (2001).
[23] D. Layzer, Astrophysical Journal **206**, 559 (1976).
[24] E. Lieb and J. Yngvason, Physics Reports **310,** 1 (1999).
[25] Lucretius, *On the Nature of Things*, trans. by William E. Leonard, e-text version, Project Gutenberg, Urbana, 1997.
[26] R. Penrose, *The Emperor's New Mind*, Oxford University Press, Oxford, 1989.
[27] S. Perlmutter et al., Astrophysical Journal **517**, 565 (1999).
[28] H. Price, *Time's Arrow and Archimedes' Point*, Oxford University Press, Oxford, 1996.
[29] H. Price, British Journal for the Philosophy of Science **53**, 83 (2002).
[30] M. J. Rees, The Observatory **89,** 193 (1969).
[31] A. Riess *et al.*, Astrophysical Journal **560**, 49 (2001).
[32] O. Spengler, *Decline of the West*, edition by Alfred A. Knopf Publisher, New York, 1996.
[33] V. S. Steckline, American Journal of Physics **51**, 894 (1983).
[34] F. J. Tipler, The Observatory **102,** 36 (1982).
[35] F. J. Tipler, International Journal for Theoretical Physics **25**, 617 (1986).
[36] M. Tucci, A. Contaldo, and S. Bonometto, Astrophysical Journal **562**, 583 (2001).
[37] J. Uffink, Studies in History and Philosophy of Modern Physics **32B**, 305 (2001).
[38] J. Uffink and H. Brown, Studies in History and Philosophy of Modern Physics **32B**, 525 (2001).

8

The Generalized Second Law Implies a Quantum Singularity Theorem

Aron C. Wall

1 Introduction

It is been speculated for some time that quantum effects will remove singularities from any theory of quantum gravity [1]. Singularity resolution has been attempted in both string theory [2] and loop quantum gravity [3]. Implicit in many of these works is the idea that quantum gravity will permit spacetimes to evade the classical singularity theorems of general relativity, and thus permit continuation past the would-be singularity. The classical singularity theorems all assume certain positivity conditions on the stress-energy tensor. However, all such conditions can be violated locally in quantum field theory. One might therefore suppose that in the highly quantum regions near a Big Bang or black hole singularity, temporary doses of negative energy might induce a bounce, avoiding the singularity [4].

The question thus arises whether there is a quantum mechanical generalization of any of the singularity theorems, which would make singularities inevitable even in quantum situations. Such a singularity theorem would have to have some assumption used in place of an energy condition which is valid in quantum situations. In this article the (fine-grained) generalized second law (GSL) of horizon thermodynamics will be proposed as a substitute. Since the GSL is widely believed to hold as a consequence of the statistical mechanical properties of quantum gravitational degrees of freedom [5], it is a good candidate for a physical law likely to hold even in a full theory of quantum gravity.

Penrose's singularity theorem [6] applies to classical general relativity coupled to matter obeying the null energy condition

$$T_{ab}k^a k^b \geq 0, \tag{1}$$

where k^a is any null vector. It says that on any globally hyperbolic spacetime with a noncompact Cauchy surface Σ, if there is a "trapped surface" T on Σ such that the outward-going null surface generated by T is contracting, then the spacetime cannot be null geodesically complete. The proof uses the Raychaudhuri equation to show that the null surface generated by T must have conjugate points, but this is incompatible

with the spacetime continuing any further. The assumption that Σ is noncompact is required to prevent the lightrays from simply intersecting one another outside of *T*. The assumption of global hyperbolicity is required because otherwise an initially noncompact universe can evolve into a compact universe as time passes (and in fact, there are non-globally hyperbolic black hole spacetimes which satisfy all other conditions of the theorem but have a Cauchy horizon instead of singularities [6]).

Penrose's theorem can be used to show that black holes must have singularities. By reversing the time orientation it can also show that if the universe is spatially infinite, it must have had a Big Bang singularity somewhere [6]. A further consequence is that there are no traversable wormholes [7], and that it is impossible to create an inflationary region in a laboratory without any initial singularities [8]. Analogous results show that the null energy condition precludes time machines [9] and superluminal communication [10, 11, 12, 13], and requires that all asymptotically flat or AdS spacetimes to have positive ADM mass [14, 12, 13].

However, none of these results applies to quantum mechanical systems because all such systems violate the null energy condition (1) [15]. There are also otherwise reasonable classical theories that violate the null energy condition [16]. Since negative mass objects probably imply that the vacuum is unstable, and time machines (and probably also wormholes) would spell trouble for causality [9], there ought to be some physical principle in the theory which prevents them from occurring. This principle, unlike the null energy condition, would have to be true in quantum mechanical situations—ideally, in some complete theory of quantum gravity.

As a step in this direction, Graham and Olum [17] pointed out that the self-consistent semiclassical averaged null energy condition on achronal[1] null geodesics was sufficient to rule out time machines, traversable wormholes, and negative energies.[2] Then [18] showed that to first order in $\hbar$, the GSL implies the condition of Graham and Olum. This means that in any situation where quantum effects are weak, the no-go results described in [17] will follow.

In this article it will be shown more generally that the fine-grained GSL can be used to prove the inevitability of singularities, and the absence of traversable wormholes, warp drives between points at null infinity, time machines, and negative mass objects, even in quantum mechanical situations. It will also be shown that no baby universes forming inside of black holes can be viable (in the sense of eventually becoming causally independent of the mother universe), and that it is impossible to restart inflation in the interior of an asymptotically flat (or AdS) spacetime.

Horizon thermodynamics also severely constrains models in an FRW-like cosmology originates out of some other pre-Big Bang universe. In this context there are interesting constraints coming from the coarse-grained GSL, as well as the time-reverse of the fine-grained GSL.

The various results will be shown first in the context of semiclassical gravity, in which one assumes that quantum effects are small, and can be controlled with an $\hbar$ expansion. However, this semiclassical approximation is used only in low curvature regions far from the singularity/pathology in question. In the high curvature region, other than the GSL itself, the results only require that basic concepts such as causality, predictivity, and topological compactness continue to have meaning in the theory of

quantum gravity. Thus it is reasonable to believe that the results will hold in a complete theory of quantum gravity. If this conclusion is correct, we must either give up horizon thermodynamics, or else several of the dreams of science fiction fans.

Since there is no well-understood nonperturbative theory of full quantum gravity (let alone an experimentally-tested one), it is of course impossible to speak with total confidence regarding the extension of these results to this regime, which is likely of importance near singularities. It may be that the concepts used to define the GSL apply only to semiclassical or perturbative gravity, not to the microscopic theory. So a conservative interpretation of the restrictions is simply that such-and-such cannot occur except by means of nonperturbative quantum gravity effects.[3] However, I will argue for a more expansive interpretation.

The plan of this paper is as follows: section 2 discusses the second law of thermodynamics, both in its ordinary and in its generalized form, with specific care given to the distinction between fine-grained and coarse-grained entropy, and a discussion of in which senses the second law does or does not rely on a well-defined arrow of time. Section 3 proves some theorems about the generalized entropy which will be used later, the most important of which is Theorem 4 which generalizes the notion of a trapped surface to quantum spacetimes. (The casual reader may wish to skim this section). Section 4 applies the GSL to obtain the various results described in the abstract. Their dependency relationships are shown in Fig. 8.1.

Up to this point, I will make free use of semiclassical notions of spacetime, even though such concepts are not valid in the full quantum gravity regime. In section 5, I will argue that the results likely apply even when quantum gravity effects are taken into consideration. Finally, the Appendix proves a theorem used in section 2.1 to help prove the ordinary second law of thermodynamics.

2 The Second Law of Thermodynamics

2.1 The Ordinary Second Law

The Ordinary Second Law (OSL) of thermodynamics states that the total entropy of a closed system cannot decrease as time passes. For the purposes of this article I will use as my definition of the entropy of a state ρ the von Neumann entropy

$$S = -\mathrm{tr}\left(\rho \ln \rho\right). \tag{2}$$

This is the analogue for quantum states of the classical Gibbs entropy which is defined as

$$S = -\sum_i p_i \ln p_i \,, \tag{3}$$

where p_i is the probability to be in the state i. Eq. (2) reduces to Eq. (3) whenever the density matrix ρ is diagonalized.

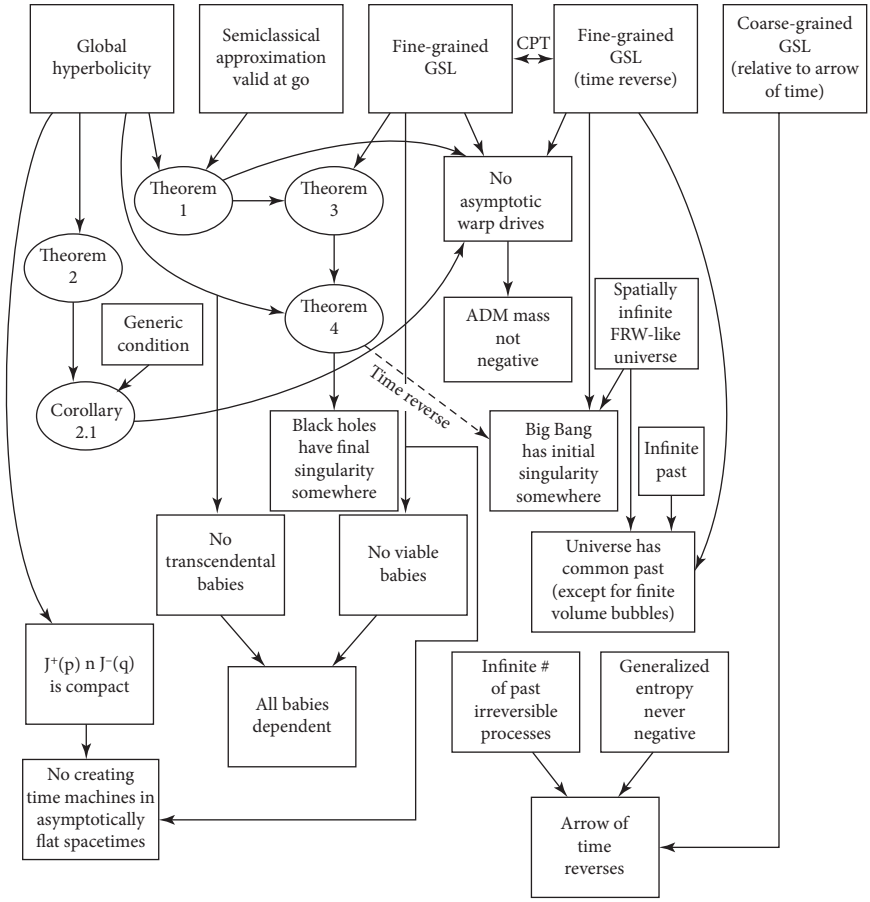

Figure 8.1 The logical dependencies of the main hypotheses, theorems, and results in this article. If a proposition has one or more arrows pointing to it, then the conjunction of *all* propositions pointing to it is used in the proof of that proposition. g_0 is a particular point in the spacetime where the semiclassical approximation must be valid; its location depends on the particular result being considered (cf. section 4). The no-time-machines result uses the assumption that $J^+(p) \cap J^-(q)$ is compact for any points p and q; since this is weaker than global hyperbolicity it is shown as following from it, although going directly from global hyperbolicity to no-time-machines is of course trivial.

In order to interpret the meaning of the von Neumann entropy, one needs to know whether ρ is interpreted in a fine-grained sense as the complete information about a state, or in a coarse-grained sense as the information available to an observer. In the fine-grained picture, closed quantum systems evolve by unitary evolution as time passes:

$$\rho(t) = U(t)\rho(t_0)U\ (t)\,. \tag{4}$$

Since unitary evolution does not affect the probability eigenvalues of ρ, this implies that

$$S\big(\rho(t)\big) = S\big(\rho(t_0)\big). \tag{5}$$

The good news is that we have just proven the OSL because the entropy cannot be decreased. The bad news is that although the entropy cannot decrease, it cannot increase either, so that the time-reverse of the second law, which I will denote as $\overline{\text{OSL}}$ also holds. This is because the entropy is a measure of the uncertainty in ρ, but the information content in ρ is just the same as the uncertainty in the initial conditions.

In order to see entropy increasing, we have to move to the coarse-grained picture. This picture can be motivated by taking a more realistic view of our ability to calculate the state at a time t from the initial conditions. If you are a Laplace's Calculator with a full understanding of the laws of nature and an infinite calculational capacity, you might well use Eq. (4) to determine $\rho(t)$. But the universe contains many complex systems for which I at any rate would be unable to specify U. Given my ignorance of the exact dynamics of the universe, I cannot fully know what $\rho(t)$ is, even if I know the initial state $\rho(t_0)$. The best I can do is rely on the things I do know about the dynamics to produce my best guess as to what I think the state is—call this $\tilde{\rho}(t)$. Since I know that the dynamics are unitary, but I do not know the exact unitary laws of physics, I ought to be able to model my ignorance as a probability distribution over the space of possible unitary processes U. This implies that I must be more uncertain about the universe at time t than a Laplace's Calculator would be, so

$$S\big(\tilde{\rho}(t)\big) \geq S\big(\rho(t)\big) = S\big(\rho(t_0)\big), \quad t > t_0. \tag{6}$$

(This equation follows from the fact that entropy is a convex function, which was first proven for quantum systems by Delbrück and Molière [19]. The Appendix provides another proof following the methods of Uhlmann [20].)

Eq. (6) shows that the entropy at any time must be greater than the entropy of the initial state. This does not quite prove the OSL, because it is not yet shown whether

$$S\big(\tilde{\rho}(t_2)\big) \geq S\big(\tilde{\rho}(t_1)\big), \quad t_2 > t_1 > t_0. \tag{7}$$

It might be, for example, that the history of the universe from t_0 to t_1 consists of some complex, calculationally intractable process, but the history of the universe from t_1 to t_2 consists of an exact reverse of that process. Then the entropy would increase at first and then decrease again later. In order to get the OSL, we need to know that this sort of thing does not happen in the real world, i.e. the complex processes which lead us to approximate the state of the universe with $\tilde{\rho}$ really are irreversible processes. Another way of putting this is that once we evolve from $\rho(t_0)$ to $\tilde{\rho}(t_1)$, it must be possible to use $\tilde{\rho}(t_1)$ as a new initial condition for purposes of determining $\tilde{\rho}(t_2)$.

A trade-off has been made here. Although the coarse-grained OSL seems to predict that the entropy will increase rather than just remain constant, by virtue of

the time-reversal symmetry of the laws of physics,[4] this is only possible if there is a time-asymmetrical assumption hidden in the proof. And there is such an assumption, embedded in the initial condition $\rho(t_0)$. In order to get a nontrivial entropy increase, $\rho(t_0)$ must have less than the maximum possible entropy. In other words, the universe has to have started out with low entropy.[5] This means that the coarse-grained OSL only holds in some states (those which really did have a low-entropy beginning), unlike the fine-grained OSL and $\overline{\text{OSL}}$ which hold in every state.

The underlying time-symmetry of the argument can be illustrated by imagining that the universe had an infinite past before the "initial" condition. Then for times $t_{-\infty} < t < t_0$, the same arguments given above show that entropy must be decreasing, so that the thermodynamic arrow of time is reversed. From a thermodynamic point of view one might prefer to describe such a universe as "beginning" at t_0 and then evolving "forwards" in time in both directions from t_0.

In order to have entropy increase for all time even with an infinite past, one might try to impose the initial condition at $t_{-\infty}$ instead of t_0. However, one would then expect that the universe would have already arrived at thermal equilibrium by any finite time t—assuming that there are an infinite number of potentially entropy producing processes before the time t.

This conclusion might be evaded if the laws of physics permit the total entropy of the universe to increase indefinitely without ever coming to equilibrium. This might actually be the case in theories of gravity similar to general relativity [21]. In an expanding universe the total volume of space can grow without limit. In classical general relativity, black holes have zero temperature, and can therefore store an arbitrarily large amount of entropy using an arbitrarily small amount of energy. Even semiclassically, it is possible for a thermal black hole to absorb an arbitrarily large amount of entropy, if it is critically illuminated for a long period of time [22].[6] Then in an infinitely large cosmos, one might have an entropy which is both infinite and increasing at all times.

Quantum modifications to general relativity may lead to even more exotic possibilities for cosmologies in an eternal-steady state entropy increase. One proposal is that each universe can spawn new universes [23, 24], each of which might continue to increase in entropy without any violation of the OSL. It has even been suggested that baby universes may have slightly different laws of physics leading to Darwinian adaptation of universes [25]. Or the universe might go through a series of cycles of de Sitter expansion and thus grow its volume and entropy without limit, as in the ekpyrotic model [26]. Are such pictures possible? In order to answer that question, we will now explore the generalization of thermodynamics to gravitational systems.

2.2 The Generalized Second Law

One comparatively simple modification which must be made to the laws of thermodynamics when taking gravity into account, is that there is no longer an absolute notion of time; there are many equally good "t" coordinates that can be used.

Since the OSL above was formulated in terms of a "t" coordinate, it is necessary to modify the OSL by considering evolution from an arbitrary complete spatial slice Σ to a complete spatial slice Σ' which is nowhere to the past of Σ. One can then formulate the OSL as the statement that the von Neumann entropy of Σ' must be at least as great as the entropy of Σ.

But there is a more profound modification to thermodynamics which arises for quantum fields in gravitational settings, which is that the laws seem to still apply in the case of certain *open* systems. One example of such a system is the exterior of a black hole. In this case there are gedankenexperiments [27] and partial proofs [28, 29] which show that the generalized entropy, defined as follows, is nondecreasing with time:

$$S_{\text{gen}} = S_{\text{H}} + S_{\text{out}} \,. \tag{8}$$

Here S_{out} is the entropy of everything outside the black hole and S_{H} is the entropy of the horizon itself, each defined on the same spatial slice Σ. S_{H} depends on the gravitational Lagrangian [30], and for general relativity takes the form

$$S_{\text{H}} = \frac{A}{4\hbar G}.^{7} \tag{9}$$

Naïvely one might have thought that one could make the entropy outside of a black hole go down by simply throwing entropy across the event horizon. But such entropy tends to be accompanied by energy, which in turn increases the mass of the black hole and correspondingly S_{H}. Similarly, Hawking radiation reduces the size of the black hole but the decrease in S_{H} is compensated for by the increase of entropy outside of the black hole [31].

There are some nuances in the definition of S_{gen}. S_{out} includes a divergent contribution coming from the short-distance entanglement entropy of quantum fields near the horizon. This divergence is quadratic with respect to a UV length cutoff. Thus S_{out}, defined as the von Neumann entropy (2) is formally infinite and requires renormalization. A second issue is that in perturbative quantum gravity, renormalization should lead to higher-order terms in the Lagrangian which renormalize Newton's constant G and also add higher order curvature terms, leading to cutoff-dependent corrections to S_{H}. The good news is that these two problems seem to cancel each other out—i.e. the divergence in S_{out} can be absorbed into the coupling constants that appear in S_{H}. This has been shown to one loop order for certain scalar and spinor theories [32], but there is an additional term appearing in the horizon entropy for gauge theories which is still not well understood [33].

Another important question is whether the GSL applies to any other horizons besides black hole event horizons. The answer seems to be yes: horizon thermodynamics seems to apply to de Sitter and Rindler horizons as well [34, 29]. However, the GSL does not hold on all null surfaces [28]; for example the past lightcone of a point has decreasing area classically, leading to a $O(\hbar^{-1})$ decrease in the generalized entropy, but the increase

in S_{out} due to quantum effects is of order $O(\hbar^0)$ and therefore cannot balance it out. The GSL also seems to be violated semiclassically on apparent horizons [35].

Following Jacobson and Parentani [34], I will assume that the GSL applies to the "future causal horizon" of any future-infinite timelike worldline W_{fut} (an "observer"). This causal horizon is defined as $H_{fut} = \partial I^-(W_{fut})$, the boundary of the past of the observer (Fig. 8.2).

Given any two complete spatial slices Σ and Σ' with the latter nowhere to the past of the former, the GSL then says that:

$$S_{gen}\left(\Sigma' \cap I^-\left(W_{fut}\right)\right) \geq S_{gen}\left(\Sigma \cap I^-\left(W_{fut}\right)\right), \tag{10}$$

where here the "outside" of the horizon used to define S_{gen} is taken to be the side on which the observer is, even for cases like de Sitter space where the observer is enclosed by the horizon (see Fig. 8.2 for examples of pairs of slices for which the generalized entropy increases.)

Note that by continuity W_{fut} may also be taken to be a lightlike ray whose affine parameter is infinite to the future, since there exist accelerating timelike worldlines which asymptotically approach any lightlike ray. In this case W_{fut} may lie on its own horizon.

Just like the OSL, the GSL comes in two versions depending on whether we choose the fine-grained or coarse-grained definition of the state ρ used to compute the entropy. In the case of the fine-grained GSL, there can still be a nontrivial entropy increase due to the fact that information can fall across the horizon between Σ and Σ'. (Another way of saying this is that even in the fine-grained picture we are still coarse-graining over all the information inside the horizon, a fully objective form of coarse-graining [36]). Since this is the only way entropy can change in the fine-grained picture, it follows that the only part of a spatial slice Σ that matters is where it crosses the horizon.

By analogy to the fine-grained OSL, the fine-grained GSL ought to hold for every state of the universe, without needing to impose any initial condition. (This can be

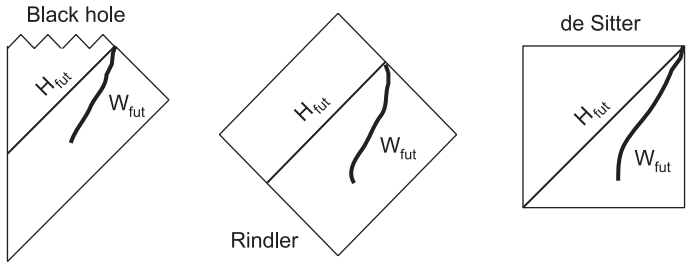

Figure 8.2 Black hole horizons, Rindler horizons, and de Sitter horizons are all special cases of "causal horizons". The future causal horizon H_{fut} is defined as the part of the spacetime which is causally visible to some future-infinite timelike worldline W_{fut}, shown as a thick line. The GSL declares that the entropy is increasing with time on complete spatial slices outside of H_{fut} (shown as red lines).

explicitly checked for many of the existing proofs of the GSL in particular regimes [28, 29].) And if the GSL is true in all states, its time-reverse must also be true in all states [18].[8]

The $\overline{\text{GSL}}$ states that for any past-infinite worldline W_{past}, the past horizon $H_{\text{past}} = \partial I^+(W_{\text{past}})$ cannot *increase* as time passes:

$$S_{\text{gen}}\left(\Sigma' \bigcap I^+\left(W_{\text{past}}\right)\right) \leq S_{\text{gen}}\left(\Sigma \bigcap I^+\left(W_{\text{past}}\right)\right). \tag{11}$$

Of course, it does depend on the initial conditions whether there *are* any past-infinite worldlines to which the $\overline{\text{GSL}}$ might be applied. If there are none then the $\overline{\text{GSL}}$ is trivially true, although it may still be useful in astrophysical settings in which the spacetime may be treated as asymptotically flat and there are approximate W_{past}'s.

On the other hand, the coarse-grained GSL would also take into account any entropy production of the matter outside the event horizon. This has the advantage of treating ordinary thermodynamic processes on the same footing as the horizon thermodynamics, but has the disadvantage that the truth of the GSL must now depend on the existence of a low-entropy initial condition. In particular the spacetime volume between Σ and Σ' must have its thermodynamic arrow of time pointing to the future.

Except for the discussion of the arrow of time in a past-infinite universe (section 4.2), the results of this article will use only the fine-grained version of the GSL.

The GSL as I have defined it has been proven for semiclassical rapidly-changing perturbations to stationary horizons, for free fields and/or Rindler horizons [29]. Proofs are also available for classical spacetimes, and for semiclassical quasi-steady processes [28]. Whether the GSL holds in a full theory of quantum gravity is obviously less certain (cf. section 5). And in the case of higher-curvature corrections to Einstein gravity, it is not yet known whether even a classical second law holds [37], except in the special case of $f(R)$ gravity [38].

3 Generalized Thermodynamics Theorems

3.1 Monotonicity properties of the generalized entropy

Suppose we have a region of spacetime R which is well described by semiclassical gravity. "Semiclassical" is a term with multiple meanings [28], but I will take it to mean the following:

Semiclassical Expansion: A region R will be said to be semiclassical if its physics can be accurately described by a finite number of terms in an expansion controlled by $\hbar G/\lambda^2$, where λ is the length scale of whatever quantum fields are relevant to the problem. This is a bootstrapping procedure in which we start with (i) a fixed classical background metric, (ii) quantize matter fields and/or linearized gravitons on this background, (iii) allow these fields to infinitesimally perturb the background due to nonlinear gravitational effects, (iv) allow that perturbation to the background to affect

the matter fields again, etc. For simplicity we will hold G and λ fixed, and write the terms of the expansion with respect to $\hbar$.[9]

For most purposes involving gravitational thermodynamics, it is sufficient to stop at step (iii), that step being needed only to calculate changes in the Bekenstein-Hawking entropy (which has an $\hbar^{-1}$ in the denominator).

We assume (without proof) that this procedure can be made well-defined using perturbative quantization of gravitons, which despite its nonrenormalizability should be valid as an effective field theory when treated using an ultraviolet cutoff much less than the Planck scale [39]. We assume that at finite orders in $\hbar$, this can be treated as if it were an ordinary quantum field theory with unitary evolution between Cauchy surfaces. When $\hbar$ is infinitesimal the gauge symmetries of the graviton correspond to *infinitesimal* diffeomorphisms. This indicates that any observable O of order $\hbar^n$ can be localized on the background spacetime up to terms which are higher order in $\hbar$.

In this semiclassical context, the generalized entropy will be assumed to take the following form:

Generalized Entropy: The generalized entropy of any codimension 2 surface will be assumed to take the form

$$S_{\text{gen}} = \frac{A}{4\hbar G} + Q + S_{\text{out}}\,, \tag{12}$$

where A is the expectation value of its area and S_{out} is the von Neumann entropy of the region spatially exterior to it, $G > 0$ is the value of Newton's constant at the renormalization scale, and the correction Q to the Bekenstein-Hawking entropy is assumed to be subleading in $\hbar$ (or some other small parameter such as string length). This is natural if Q comes from radiative corrections, as described in section 2.2.

We assume that divergences in S_{out} can be regulated using some ultraviolet regulator such as the mutual information [40]. This regulator must cut off the entanglement entropy at distance scales less than some ε, much smaller than the length scale λ of the quantum fields (so as to capture all the convergent physics) yet larger than L_{Planck} so as to avoid the quantum gravity regime.

Because fine-grained entropy is conserved, all Cauchy surfaces of the exterior should have the same value of S_{out}.

In the semiclassical regime one can show the following useful theorem about the increase of the fine-grained generalized entropy when comparing two null surfaces:

Theorem 1: Let N and M each be future null surfaces of codimension 1, each of which divides spacetime into two regions, an "interior" Int and an "exterior" Ext. Let M be either within or on N everywhere (i.e. $M \cap \text{Ext}(N)$ is empty). (The location of the null surfaces might in general depend on the state of the fields.)

Let there be a null geodesic g which lies on both N and M, and a time slice Σ which intersects g at g_0. Assume that in some neighborhood of g_0, the spacetime is semiclassical, and N and M are both smooth.[10] Very close to g_0, the null surfaces N and M will nearly coincide, but they may be separated by a small proper distance f.

For any neighborhood in the vicinity of g_0, there exists a way to evolve the time slice Σ forwards in time in that neighborhood to a new slice Σ', such that the generalized entropy increases faster on M than on N:

$$\Delta S_{\text{gen}}\left(\Sigma \cap \text{Ext}\left(M\right)\right) - \Delta S_{\text{gen}}\left(\Sigma \cap \text{Ext}\left(N\right)\right) \geq 0\,. \tag{13}$$

where Δ indicates the change in a quantity when evolving from Σ by Σ'.[11] Theorem 1 will be proven using a series of three Lemmas.

Lemma A: At the point g_0, the surface M is expanding at least as fast as N is. See Fig. 8.3.

Proof: Since N and M coincide and are smooth at the point g_0, and M cannot cross over from Int(N) to Ext(N), N and M must share the same tangent plane. The null extrinsic curvature of one of the null surfaces, e.g. N is defined as:

$$B_{ab} = h_a^c h_{bd} \nabla_c k^d \tag{14}$$

where h_{ab} is the pullback of the metric tensor onto the codimension 2 surface $\Sigma \cap N$, and k^a is a (future-oriented) null vector pointing in the direction of the null generators on N.

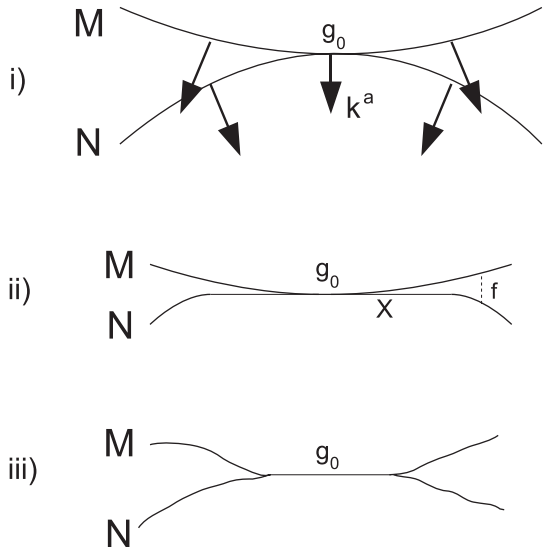

Figure 8.3 Two null surfaces N and M are pictured as they appear at one time, on the slice Σ. N is nowhere inside of M, and coincides with M at g_0. (i) the generating null vectors k^a, projected onto the surface Σ, must be normal to the null surfaces. Because M can only bend inwards relative to N at g_0, it is expanding faster than N there (Lemma A). (ii) f is the proper distance between the two null surfaces N and M, viewed as a function of M. Near the point g_0, f is very gently sloped, and thus points on N and M may be identified. Integration of $\nabla^2 f$ shows that it is always possible to find a point X near g_0 at which M is expanding faster than N, unless (iii) the surfaces coincide exactly in a neighborhood of g_0 (Lemma B).

The null extrinsic curvature measures the change in the geometry of the null surface as it moves in the k^a direction. There are two contributions. The first is a temporal component which arises when the slice $\sum$ itself has extrinsic curvature, but this may be disregarded because it is the same for both N and M. The second is a spatial component proportional to the extrinsic curvature K_{ab} of $\sum \cap N$ in $\sum$ (with a normalization factor coming from the size of the projection of k^a onto the slice $\sum$). For any vector v_a and point x, the extrinsic curvature component $K_{ab}v^a v^b$ (x) measures how much the surface N curves away from its tangent plane, to second order, as one travels away from x in the direction of v^a. A positive value means that it curves away from the direction of motion of the null surface, and a negative value means that it curves towards the direction of motion.

The fact that M is inside of N now places constraints on the extrinsic curvature of N and M at g_0. Since N is outside of M, N must bend outwards by at least as much as M does. Hence:

$$B^{(M)}_{ab} \upsilon^a \upsilon^b \geq B^{(N)}_{ab} \upsilon^a \upsilon^b \,. \tag{15}$$

The expansion of a null surface is related to the null extrinsic curvature as follows:

$$\theta \equiv \frac{1}{A}\frac{dA}{d\lambda} = B_{ab} h^{ab} \,, \tag{16}$$

where A is the infinitesimal area near a generator, and λ is an affine parameter satisfying $\lambda_{,a} k^a = 1$. Eq. (15) then requires that in the neighborhood of g_0,

$$\theta^{(M)} \geq \theta^{(N)} \,. \tag{17}$$

Q.E.D.

If the strict inequality $\theta^{(M)} > \theta^{(N)}$ holds, then by continuity inequality (17) also holds in a neighborhood of the point g_0. In the classical limit S_{out} can be neglected, so θ gives the change of entropy. In this special case, Theorem 1 follows. In the saturated case where $\theta^{(M)} = \theta^{(N)}$, Lemma A is not enough. In order to prove the classical version of Theorem 1, it is necessary to move a small distance away from the point g_0:

Lemma B: In any small neighborhood of g_0, either there is a point X at which $\theta^{(M)} > \theta^{(N)}$, or else M and N coincide everywhere in that neighborhood. In the former case, the area increases faster on M than N when $\sum$ is pushed forwards in time sufficiently close to the point X; in the latter case, the area increase is the same for M and N in the whole neighborhood. Either way, Theorem 1 holds classically.

Comment 1.1: Lemma A is a special case of Lemma B, and was proven separately for pedagogical reasons.

Proof: On the spatial slice $\sum$, let the shortest proper distance between the surfaces M and N be given by a function $f(M)$. Since the tangent planes of N and M coincide at g_0, f vanishes to zeroth and first order as one moves away from g_0. Since f is only nonzero

at second order and higher, in a neighborhood of lengthscale e, $\varepsilon^2 \ll \varepsilon$. Because the distance between M and N is in this sense small, it is possible to identify points on N and M, permitting the function to be defined on either of the two null surfaces: $f(N) = f(M)$.

This identification of points on N and M also allows the null generating vectors k^a to be compared on N and M (Fig. 8.3). When the k^a of N or M is projected onto $\sum$, it must be normal to that surface, because a lightfront always travels in the direction perpendicular the front itself. So $k^a(\sum) = cn^a$, where n^a is an outward pointing normal vector and $c > 0$ is an arbitrary constant depending on the normalization of the affine parameter λ on N and M. In order to compare the k^a vectors, λ will be chosen so that $c = 1$ everywhere on N and M.

For small ∇f this can be used to find the difference between k^a on N and M.

$$\Delta k^a = k^{a(M)} - k^{a(N)} = \nabla^a f + \mathrm{O}\left(\left(\nabla f\right)^2\right), \tag{18}$$

where up to the higher order terms, Δk^a lies on the $D-2$ dimensional surface $M \cap \sum$ (or $N \cap \sum$). The extrinsic curvature difference can now be calculated from Eq. (14):

$$\Delta B_{ab} = B_{ab}^{(M)} - B_{ab}^{(N)} = \nabla_a \nabla_b f \,, \tag{19}$$

where the covariant derivatives are intrinsic to the surface $\sum \cap N$. Together with Eq. (16) this implies

$$\Delta\theta = \theta^{(M)} - \theta^{(N)} = \nabla^2 f, \tag{20}$$

which is a total derivative. Let M (or N) be labelled by an r coordinate representing the proper distance from g_0, and let $d\sigma$ be the volume element on the $(D-3)$ dimensional space of constant r on $M \cap \sum$.

Let us define a Green's function $G(y)$ on the ball of points y with $r < R$, to be the solution to these equations:

$$-\nabla^2 G\left(y\right) = \delta^{D-2}\left(y\right); \quad G\big|_{r=R} = 0 \,. \tag{21}$$

For a sufficiently small R, the metric h^{ab} is very close to being a flat Euclidean metric, so that $G \propto (r^{D-4} - R^{D-4})/(D-4)$ (or $\ln(R/r)$ in $D = 4$). In any dimension, $G(y) > 0$ for $r < R$, and thus $\partial_r G\big|_{r=R} < 0$. For sufficiently small R these inequalities must continue to hold if the metric is slightly deformed by nonzero curvature. One can now use G to integrate $\Delta\theta$ on the codimension 2 ball B:

$$\int_B G\Delta\theta d^{D-2}y = \int_B G\nabla^2 f d^{D-2}y = -\int_{\partial B} f\partial_r G d\sigma \geq 0 \,. \tag{22}$$

where we have integrated by parts twice and used the fact that $f(0) = 0$.

Now either (i) $f = 0$ in a neighborhood of x, or else (ii) one can find arbitrarily small values of R such that the right-hand side of Eq. (22) is strictly positive, in which case $\Delta\theta$ must also be positive for at least some points arbitrarily close to x. Q.E.D.

The subject of the third lemma is the outside entropy term S_{out}, about which nothing has yet been shown.

Lemma C: If the two surfaces N and M coincide in a neighborhood of g_0, and Σ is evolved forwards in time to Σ' in this neighborhood, the entropy S_{out} is increasing faster on M than on N.

Proof: There is an information theoretical quantity called the mutual information $I(B, C)$, defined for any two disjoint systems B and C, which measures the amount by which the entropy fails to be additive:

$$I(B,C) = S(B) + S(C) - S(B \cup C). \tag{23}$$

The mutual information measures the amount of entanglement between the systems B and C. For all quantum mechanical systems, this quantity is monotonically increasing as one increases the size of one of the systems by adding a third system D [40]:

$$I(B, C \cup D) \geq I(B,C). \tag{24}$$

This makes sense intuitively, since one expects that the amount of entanglement between two systems can only be increased when one system is enlarged. This property can be exploited by setting:

$$B = \text{Int}(N) \cap \text{Ext}(M) \cap \Sigma, \tag{25}$$

$$C = \text{Ext}(N) \cap \Sigma', \tag{26}$$

$$D = N \cap \Delta\Sigma, \tag{27}$$

where $\Delta\Sigma$ is the spacetime volume between Σ and Σ'. See Fig. 8.4.

Now by Eq. (23),

$$I(B,C) = S(B) + S(\text{Ext}(N) \cap \Sigma') - S(\text{Ext}(M) \cap \Sigma') \tag{28}$$

Similarly,

$$I(B, C \cup D) = S(B) + S(\text{Ext}(N) \cap \Sigma) - S(\text{Ext}(M) \cap \Sigma), \tag{29}$$

where the slice $C \cup D$ has evolved backwards to the surface Σ, using the fact that unitary time evolution preserves the entropy. Substituting Eq. (28) and Eq. (29) into the monotonicity Eq. (24), one obtains

$$\Delta S_{\text{out}}(M) \geq \Delta S_{\text{out}}(N), \tag{30}$$

which shows that the outside entropy is increasing as fast for *M* as for *N*. Q.E.D.

Proof of Theorem 1: In the semiclassical limit, any effect which is higher order in $\hbar$ will be dominated by any nonzero effect which is lower order in $\hbar$. Let the leading order contribution to $\theta^{(M)} - \theta^{(N)}$ be of order $\hbar^{p+1}$, which corresponds to an order $\hbar^{p}$ contribution to $\Delta S_{H}^{(M)} - \Delta S_{H}^{(N)}$, since the Bekenstein-Hawking entropy (9) has an $\hbar$ in

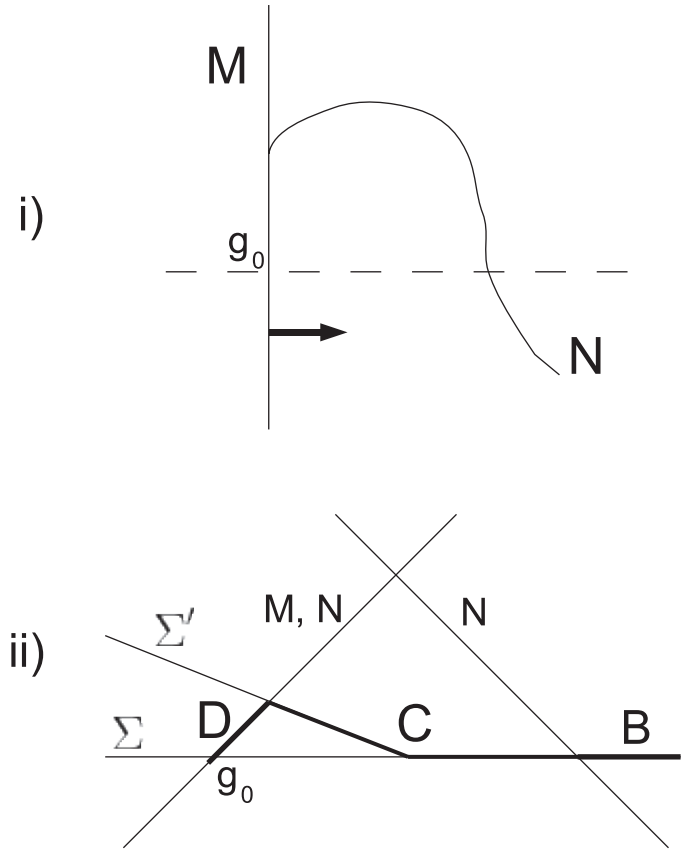

Figure 8.4 (i) Two null surfaces *M* and *N* drawn on a time slice Σ, where *N* is nowhere inside of *M*. They coincide in a neighborhood of g_0. This is the same situation as Fig. 8.3, illustrated with a different choice of *M* and *N*. The dotted line is the spatial cross-section used to make (ii) a spacetime diagram of the same situation. *N* and *M* move outwards at the speed of light. The time slice Σ is evolved forwards in time to a new time slice Σ' near g_0. All the information in $\text{Ext}(M) \cap \Sigma$ is contained in three regions: *B, C* and *D*. Removal of the region *D* can only decrease the amount of entanglement between *B* and *C*, which can be used to show that the entropy outside of *M* increases faster than the entropy outside of *N*.

the denominator. Lemma B says that at every order in $\hbar$, either N and M coincide or else $\Delta S_{\rm H}^{(M)} - \Delta S_{\rm H}^{(N)} > 0$ for an appropriate choice of $\sum$ evolution. By applying Lemma B to order $\hbar^{p+1}$, one obtains that the order $\hbar^p$ contribution to $\Delta S_{\rm H}^{(M)} - \Delta S_{\rm H}^{(N)}$ is positive. By applying Lemma B at order $\hbar^p$, one obtains that N and M coincide at order $\hbar^p$. Since Q is subleading, there is no $\hbar^p$ order contribution coming from $\Delta Q^{(M)} - \Delta Q^{(N)}$.

Let the leading order contribution to $\Delta S_{\rm out}^{(M)} - \Delta S_{\rm out}^{(N)}$ be of order $\hbar^q$. If $p \leq q$, then the area term dominates over the entropy term. If $p \geq q$, then since at this order the null surfaces coincide, Lemma *C* says that the $S_{\rm out}$ increases faster for M than N. Either way, Theorem 1 follows.

The only case not covered by the above argument is when both $p = q = +\infty$, i.e. when N and M coincide to all orders in $\hbar$. But then their generalized entropy is identical to all orders, and thus Theorem 1 is true. Q.E.D.

Corollary 1.2: At least semiclassically, one can extend the notion of a causal horizon to the boundary of the past of the *union* of any number of future-infinite timelike or lightlike worldlines. The reason is that any point on such a horizon must lie on the horizon of one of the worldlines, and then Theorem 1 shows that the GSL for the union is inherited from the GSL for that worldline.

Corollary 1.3: On the other hand, if one measures $S_{\rm out}$ in a region less than the whole exterior of a horizon, one does not always expect the entropy to increase. In particular, in the Hartle-Hawking state, the existence of nonzero entanglement will make Eq. (24) positive, which implies that any region less than the whole exterior will have decreasing entropy, as in Ref. [41].

Comment 1.4: The semiclassical assumption is unnecessary so long as M and N coincide in a neighborhood of g_0. That is because Lemma C depends only on purely information theoretical properties of $S_{\rm out}$, so it is only necessary to know that Ext(N) is a quantum subsystem of Ext(M).

Comment 1.5: With the possible exception of the no-warp-drive result in section 4.3, the results in section 4 only depend on the classical part of Theorem 1. That is because in those cases, the null surface M to which the theorem is applied already has a classical $\hbar^{-1}$ decrease in the generalized entropy, and the only thing which needs to be proven is that N also has decreasing generalized entropy.

Theorem 2: Let there be a globally hyperbolic region of spacetime R cut across by a null surface N into two regions P and F, such that information can go from P to F by falling across N, but not vice versa. Let $\sum$ and $\sum'$ be two Cauchy surfaces of R, with the latter nowhere to the past of the former. See Fig. 8.5. Then the generalized entropy of P minus the generalized entropy of F cannot increase as time passes:

$$\Delta S_{\rm gen}(P) - \Delta S_{\rm gen}(F) \leq 0 \tag{31}$$

Proof: The proof given in Ref. [18] is summarized here. Since the regions $I^-(N)$ and $I^+(N)$ share the same boundary N in the interior of R, the change in horizon entropy

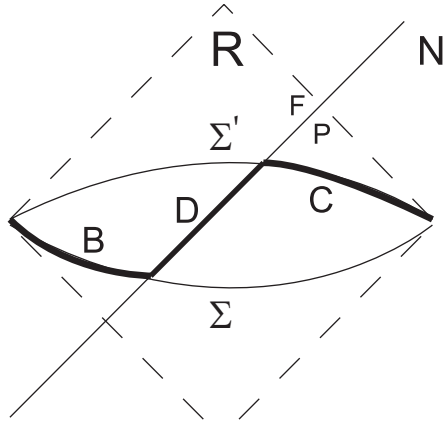

Figure 8.5 The region *R* is divided by a null surface *N* into a past region *P* and a future region *F*. A time slice Σ of *R* is evolved forwards in time to Σ'. All information in *R* is stored in the three regions *B, C,* and *D*. Weak monotonicity implies that the generalized entropy of *F* is increasing faster than the generalized entropy of *P*.

ΔS_H is the same for both of them. Furthermore any divergences in the entanglement entropy near the boundary must be the same on both sides [18]. So the only quantity which may be different is the convergent part of the S_{out} terms. In all quantum mechanical systems the entropy of three disjoint quantum systems *B, C, D,* obeys the weak monotonicity condition [42]:

$$S(B \cup D) + S(C \cup D) \geq S(B) + S(C). \tag{32}$$

Intuitively, the more a system is entangled with one quantum system, the less it can be entangled with another. Set $B = F \cap \Sigma$, $C = P \cap \Sigma'$, and $D = N \cap \Delta\Sigma$ (where $\Delta\Sigma$ is the region between Σ and Σ'). Unitary time evolution can be used to evolve the slice $B \cup D$ forwards in time onto Σ', and to evolve $C \cup D$ backwards in time onto Σ, without changing the entropy, so Eq. (32) evaluates to

$$S(F \cap \Sigma') + S(P \cap \Sigma) \geq S(F \cap \Sigma) + S(P \cap \Sigma'), \tag{33}$$

which then implies Eq. (31). Q.E.D.

Corollary 2.1: The (fine-grained) GSL and $\overline{\text{GSL}}$ cannot hold on the same null surface *N* unless they are both saturated, and weak monotonicity is also saturated. In a suitably generic state, these inequalities will not be saturated, so *N* cannot be both a past and a future horizon.

Comment 2.2: For Theorems 1 and 2, when applying the monotonicity properties (24) or (32), one may worry that the renormalization procedure needed to make S_{out}

finite will interfere with the monotonicity property. However, so long as the infinite quantities subtracted off of the entropy of a region only depend on extensive, Lorentz invariant features of the region's boundary, the divergent terms combine in such a way as to cancel out of the final result. Cf. Ref. [18] for a more detailed discussion of this point.

Comment 2.3: Because Theorem 2 follows from purely information theoretical properties of the entropy, the semiclassical approximation is not required. The only requirements are that the regions *B, C, D* be disjoint quantum systems with defined entropy, and that time evolution from $\sum$ to $\sum'$ be unitary and causal.

3.2 Quantum trapped surfaces

The GSL, as formulated in section 2.2, applies only to causal horizons. There always exist some null surfaces N for which the generalized entropy is decreasing (e.g. for $N = \partial I^-(p)$, where p is a point, the generalized entropy always decreases near p). This does not contradict the GSL so long as N is not a horizon. The GSL is thus logically equivalent to the statement that any such null surface N with decreasing entropy is not a causal horizon.

That in turn means that there does not exist any worldline W_{fut} which is infinite to the future and for which N is the boundary of the past of that observer. This can be used to show that certain null surfaces must necessarily terminate:

Theorem 3: Let N be an achronal null surface, such that g be a null generator of N, and let there be a point g_0 on g at which the fine-grained generalized entropy is decreasing. That is, there exists a spatial slice $\sum$, such that for any neighborhood around g_0, there is a way of pushing the slice forwards in time in that neighborhood to a new slice $\sum'$, so that

$$\Delta S_{\text{gen}}\left(\Sigma \bigcap \text{Ext}\left(N\right)\right) \leq 0. \tag{34}$$

Thus if N were a horizon, the fine-grained GSL would be violated for time evolution near g_0.

Suppose that the semiclassical approximation holds near g_0, while for the rest of the spacetime we assume only that it can be described by a Lorentzian manifold. Then the GSL implies that the null generator g cannot be extended infinitely on N (either because it exits N, or because spacetime is null geodesically incomplete).

Proof: If g stays on N for an infinite affine distance, then it too must be achronal. Furthermore it will have a horizon H on which the GSL is satisfied, since the GSL must also apply to the horizons of infinite null rays as discussed in section 2.2. H must lie entirely on or to the past of N, because the past of g must be a subset of the past of N. Since g is lightlike and achronal, g must be a generator of H as well as N. Then Theorem 1 implies that the generalized entropy is also decreasing on H, which would violate the GSL. Q.E.D.

If spacetime is globally hyperbolic, then a stronger result can be shown for certain surfaces.

Global Hyperbolicity: A spacetime is globally hyperbolic iff both a) there are no closed causal curves[12] and b) for any two points p and q, $J^+(p) \cap J^-(q)$ is compact.

Global hyperbolicity implies that there are Cauchy surfaces which intersect every timelike worldline exactly once. Furthermore, one can find a smooth timelike vector field, whose integral curves must intersect any Cauchy surface exactly once [6].

Quantum Trapped Surface: Let there be a connected Cauchy slice Σ, containing a compact codimension 2 surface T which divides it into two regions Ext(T) and Int(T), such that Ext(T) is noncompact. Let a null surface N be shot out from T going outwards and to the future; N may be defined more precisely as the future boundary of the domain of dependence of the exterior: $\partial^+ D[\text{Ext}(T)]$. Such a boundary is necessarily achronal. If the fine-grained generalized entropy of N is decreasing with time for each point g_0 on T (in the sense described above in Theorem 3), then T is a "quantum trapped surface".

Comment 4.1: In the classical limit, the generalized entropy becomes the area, and the definition reduces to the usual classical one: a surface T is classically trapped if the area of the surface N is decreasing everywhere at T when moving outwards to the future. By analogy to this, T is quantum trapped if it is a compact surface for which the generalized entropy is decreasing everywhere on a compact, outward-moving N, near every point g_0 of T.

Comment 4.2: The existence of a quantum trapped surface does not necessarily violate the GSL, because the GSL only applies to future horizons, and N is not necessarily a horizon. However, if a quantum trapped surface *were* a horizon, then it would violate the GSL everywhere on T.

Theorem 4: Suppose there exists a globally hyperbolic spacetime with a quantum trapped surface T, as described above. Let the semiclassical approximation be valid near T (but not necessarily elsewhere). Then the fine-grained GSL requires that the spacetime is not null geodesically complete, i.e. there is a singularity somewhere.

Proof: By Theorem 3, each of the null generator segments on N, i.e. $\overline{g} \equiv \overline{g} \cap N$, must terminate at some finite value of the affine parameter λ, because the generalized entropy is decreasing on it. From this point on, the argument is the same as the classical Penrose singularity theorem [6] which we summarize here:

Assume for contradiction that the manifold is null geodesically complete. In that case, each segment $\overline{g}$ may be extended to the future beyond N, and therefore $\overline{g}$ includes its own endpoint, as part of N. We can rescale the affine parameter so that $\lambda = 0$ at T and $\lambda = 1$ at the endpoints. This allows us to write N as the topological product $T \times [0, 1]$, except that some of the endpoints at $\lambda = 1$ may be identified with each other. Since T is compact and so is the closed line segment $[0, 1]$, it follows that N is compact.

However, global hyperbolicity prevents a noncompact spatial slice $\sum$ from evolving in time to a compact spatial slice N. To see this, choose a smooth time-like vector field t^a whose integral curves intersect $\sum$ once. Since N is achronal, the integral curves of t^a intersect N at most once. t^a can then be used to define a homeomorphism from N to part of $\sum$. Since N is compact and without boundary, it must map to a subspace of $\sum$ which is itself compact and without boundary, but this contradicts the fact that $\sum$ is connected and noncompact. Hence the spacetime must actually be null geodesically incomplete. Q.E.D.

For more detailed descriptions of the Penrose proof, see Ref. [6].

Comment 4.3: This shows that the Penrose singularity theorem can be generalized to quantum spacetimes so long as the fine-grained GSL holds. This idea that the GSL gives rise to an analogue of trapped surfaces is implicit in the "quantum Bousso bound" proposed by Strominger and Thompson [44].

Comment 5: Since by time-reversal symmetry the fine-grained $\overline{\text{GSL}}$ must be just as true as the fine-grained GSL, the time-reversals of Theorems 1-4 also hold.

4 Applications

4.1 Black holes and babies

We will now apply the fine-grained GSL to the case of black hole collapse in order to show that there must be a black hole singularity somewhere (or else a Cauchy horizon due to failure of global hyperbolicity). This requires a "quantum trapped surface" on which the GSL is being violated. For a black hole with radius $r \gg L_{Planck}$, the black hole should normally be described by an approximately classical metric. So it is sufficient to find a surface T which is classically trapped (i.e. its area is contracting before taking into account any quantum effects). This decrease of area then implies an $O(\hbar^{-1})$ decrease in the generalized entropy, which cannot be compensated for by the $O(\hbar^0)$ increase in S_{out}. Consequently the surface T is also quantum trapped, which by Theorem 4 implies that it must be null geodesically incomplete, or else not globally hyperbolic—the exact same result obtained by the Penrose singularity theorem, but now applicable to certain quantum-mechanical situations.

However, just because there is a singularity somewhere does not necessarily mean that there must be a singularity everywhere. Is it possible to avoid the singularities somehow and end up in a new universe? Let us define more carefully what we mean by a baby universe: a baby universe is a spacetime region which is 1) inside the event horizon of a black hole, 2) contained in the future of the exterior of the black hole, and 3) can last for an indefinitely long proper time as measured by at least one observer. That is, there must be able to exist a future-infinite worldline W_{fut} inside the event horizon. (Note that if the baby universe ends up in a de Sitter

type expanding phase, there may be multiple choices of W_{fut} which are separated by causal horizons.)

Proviso (1) ensures that the baby universe is distinct from the mother universe. Proviso (2) rules out other universes which are not formed wholly from our own universe, but have pasts which are causally disconnected with our own universe. In particular, the baby universe should not have come from an initial singularity of its own. For example, Schwarzschild-de Sitter would not be an example of a baby universe spacetime. Proviso (3) is necessary to distinguish a baby universe from the usual picture of a black hole interior in which everything must end on a singularity in finite proper time.[13]

Assume that the black hole is in an asymptotically flat spacetime. There are two different time coordinates which can be used to parametrize any given W_{fut}. Let there be a large stationary sphere around the black hole with a proper time coordinate t. The advanced time coordinate $v(x)$ of any point x is defined as the maximum value of t on that part of the sphere which is to the past of x. Another possible coordinate is the proper time τ of the timelike worldline W_{fut} inside of the baby universe. Baby universes can be classified by means of the monotonic relationship between the two time coordinates $v(\tau)$ as follows (see Fig. 8.6):

1. **Viable Babies:** Doctors call a fetus viable when it is capable of existing on its own without further life support from the mother's womb. Adapting this definition for baby universes, let us define a "viable baby universe" as a baby which eventually becomes capable of causally existing on its own without support from the mother. This requires that there exists *at least* one W_{fut} such that as $\tau \to +\infty$, $v \to v_{max}$ for some finite v_{max}. This is equivalent to saying that W_{fut} is in the future of a compact spacetime region.
2. **Dependent Babies:** A baby universe is dependent if it requires continued causal influences coming from the mother universe in order to remain in existence. Apart from these causal influences, the baby can only last for a finite proper time. This means that for all W_{fut}, $\tau \to +\infty$, $v \to +\infty$.
3. **Transcendental Babies:** The remaining logical possibility is that for some W_{fut}, $\tau \to \tau_{max}$ for some finite τ_{max}, $v \to +\infty$. This means that the baby universe requires an infinite period of gestation to reach a finite proper time, and then it goes on to become independent! This very odd behavior violates global hyperbolicity, because if one takes a point p outside the horizon, and a point $q \in W_{fut}$ with $\tau > \tau_{max}$), the region $I^+(p) \cap I^+(q)$ causally in between them is noncompact. Thus W_{fut} crosses a Cauchy horizon at τ_{max}. There would therefore be a failure of predictiveness across the Cauchy surface unless some new nonlocal physics principle were to come into play. An example of such a spacetime is the analytically continued Reissner-Nordström metric.

A baby universe will start out being connected to the mother universe by an umbilical cord through the black hole event horizon. However, a black hole left in a vacuum will eventually radiate away its mass until it reaches the Planck scale. What happens after that depends on one's assumptions about quantum gravity. If the black hole evaporates completely, then the connection must be broken at a finite value of v. Call

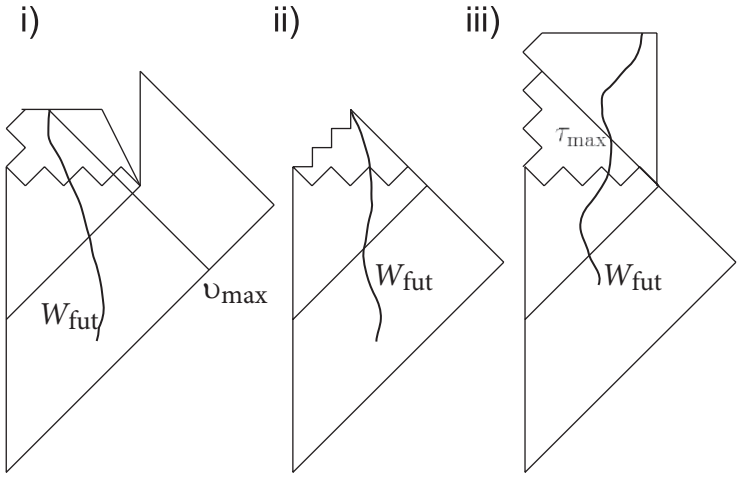

Figure 8.6 Sample Penrose diagrams for baby universes forming from a collapsed black hole. Each baby universe is shown to the future of a jagged line representing the classical singularity, and has a future infinite worldline W_{fut} falling into it. This classical singularity may be resolved in some places by quantum effects, however the GSL requires that true singularities also appear somewhere in the spacetime (also shown by jagged lines). It further places restrictions on the kinds of baby universes that are possible. (i) A viable baby universe, ruled out by the GSL. Since the baby universe ends in a de Sitter phase, several different future horizons can be selected depending on the choice of W_{fut}. In the example shown, the black hole evaporates completely, resulting in a disconnected space. However, the GSL also excludes viable babies which remain connected to the mother universe. (ii) A dependent baby universe, permitted by the GSL. The black hole cannot evaporate completely. (iii) A transcendental baby universe, ruled out by global hyperbolicity. At $\tau_{\max}$ the worldline W_{fut} crosses a Cauchy horizon. As it crosses it sees the entire history of the outside universe, infinitely blue shifted.

the resulting baby universe a *disconnected* baby. (We will assume that once the baby universe disconnects, it remains separate rather than reconnecting at a later time.) Such topology changing spacetimes are non-globally-hyperbolic [45]. This kind of violation of global hyperbolicity might well be physically reasonable though, since the loss of predictivity only occurs at a single point in the manifold. (Quantum gravity might restore predictivity, by specifying the dynamics of such topology-change points.)

The alternative is a *connected* baby, which always remains connected to the mother universe, either because there is a remnant left over from black hole evaporation, or because the black hole is illuminated by sufficient infalling matter to prevent total evaporation. It is also possible for a connected baby to be viable if e.g. it enters a de Sitter expansion phase, so that horizons form around individual future worldlines.

All viable babies—whether globally hyperbolic or not—are ruled out by the GSL, because the viability condition ensures that W_{fut} has a spatially-compact future horizon

$\partial I^-(W_{fut})$ existing outside the black hole, which separates the points which can causally influence the baby universe from the points which cannot. However, when one traces the horizon $\partial I^-(W_{fut})$ to the asymptotic region far from the black hole, then its generalized entropy must be decreasing with time. To show this, we will first consider the spherically symmetric case, and then generalize to the non-spherically symmetric case. We will assume that the semiclassical approximation is valid in the asymptotically flat region, far from the black hole.

In the spherically symmetric case, the future horizon is defined by $v = v_{max}$ (the last moment of advanced time from which a signal can reach the baby universe). Far from the black hole, the horizon is therefore a contracting sphere. Since the area of such surfaces is decreasing on the classical black hole background metric, the Bekenstein-Hawking area entropy is shrinking by an $O(\hbar^{-1})$ term. Any quantum corrections coming from S_{out} are $O(\hbar^0)$, which is of lower order in the semiclassical expansion. Consequently the generalized entropy of a future horizon is decreasing, contrary to the GSL.

If the spacetime is not spherically symmetric, then $\partial I^-(W_{fut})$ may partly lie to the past of $v = v_{max}$. However, it is still true that $\partial I^-(W_{fut})$ reaches to past null infinity, and that the asymptotic area of $\partial I^-(W_{fut}) \cap I^-$ is infinite. On the other hand, the area of a compact slice of $\partial I^-(W_{fut})$ is finite when the slice is taken at large (but finite) distance from the black hole. In order to go from infinite to finite area, there must exist some point g_0 far from the black hole where $\partial I^-(W_{fut})$ is contracting classically. But then, by the same argument as in the preceding paragraph, the generalized entropy decreases at g_0.

Therefore if the GSL is true, no such W_{fut} can exist; in other words there is no viable baby universe. This is a generalization of a theorem in classical general relativity using the null energy condition [8].

So far we have not assumed that the baby spacetime is globally hyperbolic. Global hyperbolicity would rule out the transcendental babies, as well as disconnected babies. But of the two, transcendental babies seem much more pathological because of the "infinite blueshift" of the field modes falling across the horizon at late times. So even if topology changing events are allowed, it still seems reasonable to disallow transcendental babies. That would exclude the proposed quantum tunneling process, described in Ref. [46], as well as the baby universe spacetime of Ref. [24]. (Even if transcendental babies were allowed, they would probably be very sensitive to the long-term fate of the universe, since they can only arise inside of black holes that have a finite probability of never evaporating completely.)

Thus, assuming both the GSL and global hyperbolicity, only dependent babies are permitted as a possibility. The reason why the GSL does not forbid dependent babies is that $I^-(W_{fut})$ includes the entire exterior of the black hole. Thus there is no horizon separating the points which can influence the baby from the points that cannot. However, because any babies must be dependent, the opportunities for universe creation are limited.

First of all, a dependent baby universe must always remain connected by its umbilical cord to the mother universe lest it die. But any black hole that is left to itself

will evaporate due to Hawking radiation. So unless remnants are allowed, everything inside of the black hole must be destroyed if the black hole ever stops being fed.

Secondly, a dependent baby universe cannot have a cosmology similar to our own universe, which appears to be heading into a de Sitter expanding phase. De Sitter space has a compact future horizon around any future-infinite worldline W_{fut}. Let there be a spatial slice Σ on which $\Sigma \cap J^-(W_{\text{fut}})$ is a compact region.[14] For any non-transcendental baby, v must be finite at every point in $\Sigma \cap J^-(W_{\text{fut}})$. Since v is a continuous function, by compactness, this means that v has an upper bound in the region $\Sigma \cap J^-(W_{\text{fut}})$. This means that no information can reach W_{fut} after a certain advanced time v; hence the baby universe is viable. Consequently, no dependent baby universe can end up in a de Sitter expanding phase.

So the GSL requires that any baby universes (and their progeny forever) must remain dependent on this one, and have a different cosmology than our own universe appears to. (Additional constraints on universe formation will be given by the $\overline{\text{GSL}}$ in the next section.)

Restarting Inflation. Assuming the GSL and global hyperbolicity, the same argument that rules out baby universes also implies that one cannot restart inflation in an asymptotically Minkowski space, since the compact future de Sitter horizons would violate the GSL in the Minkowski region of spacetime. In other words, any inflationary region would become a viable baby universe, and would therefore be ruled out. This corresponds to a classical result using the null energy condition [8]. It is also in agreement with the AdS/CFT argument of Ref. [47]. On the other hand, the prohibition on baby universes appears to conflict with semiclassical instanton calculations [48] of quantum tunnelling probabilities to restart inflation. However, these instanton calculations correspond to Euclidean manifolds over degenerate metrics. Their validity is controversial [49].

Traversable Wormholes. Similarly, there can be no traversable wormholes between two distant regions of spacetime, because any worldline which crossed from past null infinity of one region, to future null infinity of the other, would have to have a classically contracting future horizon in the first region. Again, this result is analogous to a classical result using the null energy condition [7].

De Sitter and Anti-de Sitter The arguments in this section can also be applied to asymptotically anti-de Sitter spacetime, and more generally to any spacetime in which large ingoing null surfaces are contracting. (Thus it does not matter that Anti-de Sitter space violates global hyperbolicity due to its boundary at spatial infinity). In the case of wormholes between two asymptotically AdS spacetimes, the prohibition of wormhole traversal is in accordance with AdS/CFT [50]. Since such spacetimes would have two disjoint conformal boundaries, there is no way that information could be causally transmitted from one CFT to the other.

In the case of de Sitter space, ingoing null surfaces are contracting only if they are sufficiently small. Therefore, the GSL only restricts baby universes and inflationary regions in de Sitter space if their horizons are at a distance scale shorter than the de Sitter radius. It may be that this places constraints on eternal inflation scenarios. However, in order to address eternal inflation it is necessary to carefully consider the role of entropy fluctuations (cf. section 5.1, footnote 19).

4.2 Big bangs and beginnings

Did the universe have a beginning in time? We have already discussed two incomplete arguments that it did: a) In section 2.1, the coarse-grained OSL was used to argue that if the thermodynamic arrow of time always points forwards, there can only be a finite amount of entropy production in our past. However, the argument failed for gravitational systems such as general relativity because of the possibility that the entropy might be able to increase without bound, permitting systems with infinite yet increasing entropy. b) In classical general relativity, one can instead use the Penrose singularity theorem to argue that if the universe is spatially infinite, there must have been an initial singularity. But this theorem uses the null energy condition, which fails for quantum fields.[15]

In this section we will use generalized thermodynamics to prove quantum analogues of both (a) and (b). Let us assume that although the early universe may have been quantum and inhomogeneous, at late times and at large-distance scales, the universe is described by some expanding classical Friedmann-Robertson-Walker (FRW) cosmology. The application of the classical Penrose theorem to the Big Bang cosmology uses the fact that a sufficiently large sphere T in an expanding FRW cosmology is an anti-trapped surface, i.e. even the inward moving null surface generated from T is expanding. When these rays are instead traced backwards in time, they are contracting and outwards moving. Assuming global hyperbolicity, at least one ray must be null geodesically incomplete, which implies a singularity.

To generalize this result to the quantum case, note that if the anti-trapped surfaces are in a semiclassical region of spacetime, the fact that T is classically anti-trapped means that it is also quantum anti-trapped, meaning that the generalized entropy of this past horizon decreases when one goes to the past. Using the fine-grained $\overline{\text{GSL}}$, it follows from Theorem 4 that the spacetime has an initial singularity, if spacetime is globally hyperbolic and space is noncompact. It does not matter if the FRW cosmology has small inhomogeneous perturbations because small perturbations cannot eliminate the anti-trapped surfaces.

Just as in the black hole case, the mere fact that there is an initial singularity, does not necessarily tell us that there are no past-infinite worldlines W_{past} which avoid the singularity. But some constraints can be placed on this possibility by assuming that there is such a W_{past} and then applying the fine-grained $\overline{\text{GSL}}$ to it. Suppose that a past horizon $\partial I^+(W_{\text{past}})$ exists in the present-day universe. Because of the expansion of the universe, such a horizon should now be a large, nearly classical object. And by the $\overline{\text{GSL}}$ it must be nonexpanding everywhere. In an expanding FRW cosmology, this is only possible if each connected component of the past horizon is compact and sufficiently small. Therefore the entirety of the infinite universe would share a common past history, except for possibly a set of bubbles each with finite spatial volume.[16]

This is in stark contrast with the standard hot Big Bang FRW cosmology, in which sufficiently distant spatial regions have never been in casual contact with one another. But it is not too different from inflationary cosmology, in which the exponential expansion of the universe causes distant regions to share a causal past. One way to

provide spacetime with the requisite property would be if there were an infinite period of inflation to the past of the infinite universe. Such a spacetime would have past-infinite worldlines, but would also be null geodesically incomplete due to the fact that an infinite inflating universe occupies only a piece of de Sitter space.

Similarly a spatially finite universe can easily have all points in its FRW phase eventually be in causal contact, and it does not even need any initial singularities to do so. An example would be a A-FRW cosmology with a bounce.

In order to analyze the thermodynamic properties of such a past-infinite model, we now invoke the coarse-grained GSL. Assuming in accordance with current observations that the universe will end up in a de Sitter-like expanding phase, there will be future-infinite worldlines W_{fut} beginning on Earth which will end up being surrounded by a compact future horizon, containing a finite amount of generalized entropy (approximately $A/4\hbar G$ of the horizon). Assuming global hyperbolicity, this future horizon must remain compact as it is taken to the past, and includes in particular everything in the past of Earth. This means that either 1) the generalized entropy inside the horizon has increased from arbitrarily negative values, or 2) there are only a finite amount of entropy producing processes in our past lightcone, or 3) the thermodynamic arrow of time is reversed somewhere in our past, so that the coarse-grained GSL does not hold.

Option (1) is not possible if the generalized entropy has a direct state counting interpretation in terms of discrete Planck-scale degrees of freedom, and in any case seems somewhat bizarre. Option (2) would involve the universe being in a near-equilibrium state for the first "half" of eternity and then for some inexplicable reason exiting this equilibrium. Option (3), although strange seeming, arises naturally if the low entropy "initial conditions" of the universe are actually imposed on some finite time slice (cf. section 2.1). In this scenario, the universe can be said to have a beginning in a thermodynamic sense even if it does not have a beginning in a geometrical sense.

Option (3) works best if the universe is spatially finite. In a spatially finite universe, the horizon of W_{fut} can, when traced backwards in time, intersect itself and disappear entirely at a time t_*. (In the case of a spatially-infinite, globally hyperbolic universe the horizon can only disappear entirely by hitting an initial singularity.) Since before t_* there is no horizon, the fine-grained generalized entropy is simply a constant, equal to the fine-grained entropy of the total universe. This explains why the fine-grained entropy does not decrease indefinitely when one goes backwards in time. But if the horizon goes back in time forever, one would run into problems with the fine-grained GSL. Since the fine-grained GSL does not depend on an arrow of time (cf. section 2.2), one would have to endorse options (1) or (2) with respect to the fine-grained generalized entropy, eliminating the benefit obtained from reversing the arrow of time.

Putting all these considerations together, if the GSL is a valid law of nature, it strongly suggests that either the universe had a finite beginning in time, or else it is spatially finite and the arrow of time was reversed previous to the Big Bang.[17] In the latter case, it could still be said that the universe had a beginning in a thermodynamic sense, because both branches of the cosmology would be to the thermodynamic future of the Big Bang.

4.3 Warp drives and negative mass objects

The fine-grained GSL can also be used to rule out certain kinds of warp drive spacetimes. We will consider spacetimes M which are asymptotically flat and globally hyperbolic, but which have in their interior some gravitational fields which are capable of bending the lightcones so as to enable superluminal travel. See Ref. [51] for discussions of such spacetimes.

Since the positions of the lightcones depends on the choice of coordinates, it is necessary to provide a diffeomorphism-invariant definition of superluminal travel. Coming up with a sensible definition is tricky, since diffeomorphisms can move around the start and finish points. In order to deal with this, this section will only consider the case of *asymptotic* warp drives, which speed up the propagation of light rays relative to the asymptotic structure of the spacetime.

The approach here is inspired by Ref. [10], which defined a lightray as being superluminal if it travels between a certain pair of 2-surfaces faster than any nearby lightray. However, instead of looking at travel between two 2-surfaces separated by a finite distance, I will consider a lightray g travelling from the past conformal boundary I^- to the future conformal boundary I^+.

Intuitively, a spacetime is a warp drive if there exists a compact region $D \in M$ such that light-speed signals which pass through D are advanced by a finite time relative to signals which "go around" in $M - D$. (See Fig. 8.7.)

More precisely, let us define a "warp drive" as a compact spacetime region $D \in M$, with the property that there exist points $L^- \in I^-$ and $L^+ \in I^+$ such that (see Fig. 8.8):

1. L^- and L^+ are achronal (i.e. they are not connected by any timelike curve),
2. There exists a null curve g travelling from L^- to L^+ passing through the region D, but
3. In the partial spacetime $M - D$, any null curves travelling from L^- to I^+ are delayed by a finite time. In other words, for any other point $p \in I^+$, if p is null separated from L^+, and if p is sufficiently close to L^+, then it is not possible to send a signal from L^+ to p.

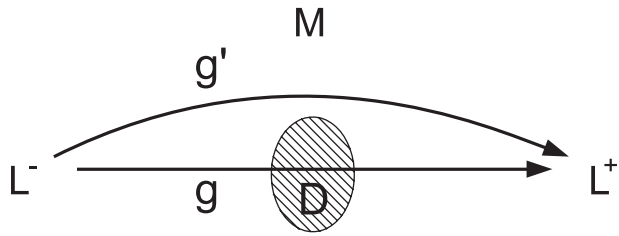

Figure 8.7 This diagram shows light signals travelling through space from an asymptotic past origin L^- to an asymptotic future destination L^+. A warp drive is defined as a compact region D of spacetime, such that some null curve g passing through D is advanced by a finite time relative to any curve g' which does not pass through g.

To summarize this definition, L^+ and L^- are lightlike separated in M, but spacelike separated on $M - D$. This indicates that null curves which travel through D are advanced by a finite time compared to those which go around D.

A few comments on this definition: First, conditions 1 and 2 implicitly require g to be an achronal null curve, which in turn implies that it is a "fastest possible" geodesic connecting I^- and I^+. The existence of a curve with maximum possible speed follows from global hyperbolicity for any warp drive space-time, since the space of causal curves between any two compact subsets of a globally hyperbolic spacetime is itself compact [52].

The points L^- and L^+ must be in diametrically opposite spatial directions, since all other points in I^+ are chronal to any L^-. This is why condition 3 above restricts attention to points in I^+ which are null separated to L^+.

One might naively that even in flat spacetime, a region D could satisfy this definition simply by blocking the shortest path between L^- and L^+, and thus forcing any lightrays connecting L^- and L^+ in $M - D$ to go around a longer way. However, because L^- and L^+ are infinitely far away, lightrays only need to bend by a small amount to get around D, leading to a delay that can be made arbitrarily small, meaning that condition 3 is not satisfied.

Finally, this definition applies only to warp drives that lead to a finite advance for lightrays travelling over an infinite distance. The result does not apply to cases where there is a speed up only over a finite distance.

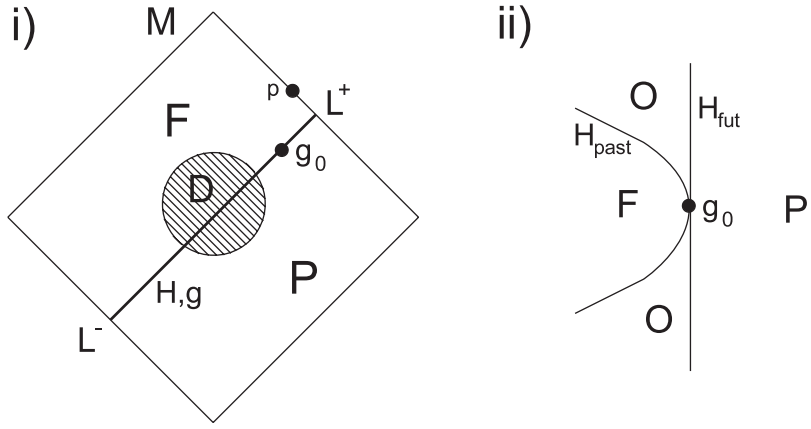

Figure 8.8 i) A spacetime diagram of the warp drive. The null curve g passes through the region D in order to connect points L^- and L^+ on the conformal boundary. There exist points p to the null future of L^+ which cannot be accessed from L^- without passing through the region D. The null curve g lies on both the future horizon $H_{\text{fut}} = \partial I^-(L^+)$ and the past horizon $H_{\text{past}} = \partial I^+(L^-)$. However, these horizons need not coincide except at g. ii) A picture of a spatial slice that cuts through the point g_0. The two horizons define three regions, F to the future of g, P to the past of g, and the spacelike separated other region O. Assuming the semiclassical approximation is valid at g_0, Theorems 1 and 2 can then be applied to show a violation of either the GSL or .

We will now show a contradiction between generalized thermodynamics and the existence of warp drives as defined above. Since there exist infinite worldlines beginning on L^- or ending on L^+, these points define Rindler-like horizons cutting through the spacetime M. These two horizons cut M into three regions, $P = I^-(L^+)$, $F = I^-(L^+)$, and the remainder $O = M - F - P$ (P and O cannot overlap or L^- and L^+ would be timelike rather than lightlike). g is required to lie in both the causal past region $J^-(L^+)$ and the causal future region $J^+(L^-)$. However, it cannot lie in the interior of either region, or there would be a timelike curve going from L^- to L^+. Consequently it must be a null geodesic lying on the boundary of these regions, on both the future horizon $H_{\text{fut}} = \partial I^-(L^+)$ and the past horizon $H_{\text{past}} = \partial I^+(L^-)$. Assuming that F obeys the GSL and P obeys the $\overline{\text{GSL}}$, it is now possible to derive a contradiction.

Although the fields may not be semiclassical in the region D, there ought to exist at least one point $g_0 \in g$ on which the semiclassical limit does hold in the asymptotically flat region away from any significant matter excitations. At this point g_0, Theorem 1 may be invoked to show that if the GSL holds on F, it also holds on $F + O$. Furthermore, Corollary 2.1 (choosing $R = M$) implies that if the GSL holds on $F + O$ and the $\overline{\text{GSL}}$ holds on P, then the state is not generic. But if there are any warp drive spacetimes, there are also generic ones, since if a lightray is advanced by a finite time by travelling through D (condition 3), an infinitesimal generic perturbation of the spacetime must preserve this property. It follows that either there are no (asymptotically flat, globally hyperbolic) warp drive spacetimes, or else the GSL or $\overline{\text{GSL}}$ are violated.

This result is a generalization of the classical theorems of Refs. [10, 11, 13].

Positive Energy Theorem. An immediate corollary is that the GSL and $\overline{\text{GSL}}$ together imply a positive energy theorem, by the principles outlined in Ref. [14]. Suppose we have an asymptotically flat spacetime containing an isolated compact object with some ADM mass. Now an object with a positive ADM mass causes lightrays passing through its asymptotic gravitational field to be delayed. For a negative mass object this Shapiro "delay" is actually an advance. Thus an object with negative mass could be used for superluminal communication, as a type of warp drive. Hence it is forbidden by the previous result.

The classical theorem [14] has an important limitation which should be noticed: it requires the spacetime to be asymptotically flat at *null* infinity, not just spacelike infinity. In other words, the spacetime must remain sufficiently stable that its mass can be probed by a lightray going from past null infinity to future null infinity. For example, there exist Kaluza-Klein spacetimes with negative ADM mass in which a "bubble of nothing" contracts and then expands, asymptotically approaching the speed of light [53]. This is not a contradiction because the bubble of nothing hits null infinity and prevents it from being asymptotically flat. Similar problems arise for bubbles of AdS space inside of a "false vacuum" with zero cosmological constant.

For asymptotically Schwarzschild solutions in 4 dimensions, one can also derive a partial converse result: any solution with positive mass is *not* a warp drive spacetime. Because the gravitational potential falls off with distance like $1/r$, the integrated time delay is logarithmically divergent, and therefore the Shapiro delay from the asymptotic gravitational field is $+\infty$. This overcomes any finite Shapiro advance coming from the

interior of the spacetime. Unfortunately, this makes the no-warp-drive result somewhat trivial in this case, since it does not rule out any asymptotically Schwarzschild solutions with positive mass.

However the result is not as trivial for $D \geq 5$ spacetimes, where the Shapiro delay for Schwarzschild is finite. Nor is it trivial for asymptotically Anti-de Sitter spacetimes, which we consider next.

Causality in Anti-de Sitter. The results above can be directly generalized to the case in which the spacetime is asymptotically AdS, by choosing L^+ and L^- to lie on the AdS boundary. The proof of the result is analogous, notwithstanding the fact that anti-de Sitter space technically violates global hyperbolicity. (The only reason why Theorem 2 needed global hyperbolicity was to ensure that the spacetime region had good causality properties. But AdS space has equally good causality properties after one imposes boundary conditions at spatial infinity.) This generalizes the classical theorems of Ref. [12, 13].

This also fits in nicely with what is known about the AdS/CFT conjecture. In order for a theory of gravity to have a field theory dual living on its boundary, it is essential that between two spatial locations A and B on the boundary, it is impossible to get from A to B any faster when travelling through the bulk, than when going around on the boundary. Otherwise, it would be possible to send signals faster than light in the CFT [54]. The GSL is a plausible physical principle enforcing this requirement.

4.4 Time machines

The final application of the fine-grained GSL will be to rule out time machines, which is again a generalization of classical results [9]. Up until now, we have assumed that spacetime is globally hyperbolic. But global hyperbolicity rules out closed timelike curves (CTC's) by definition, making any proof trivial. So in this section, we will assume instead that a) all CTC's in the spacetime are to the future of some point p, which is in turn to the future of I^-, b) for any two points p and q, $J^+(p) \cap J^+(q)$ is compact (i.e. the other component of global hyperbolicity [43]), and c) spacetime is asymptotically flat.

The goal will be to show that a CTC can never form. Since a CTC is an infinite worldline wrapped around the same points periodically, $\partial I^-(\text{CTC})$ is a future horizon.[18] Excluding the future of p, spacetime is globally hyperbolic, so there is no problem defining complete slices for the generalized entropy. By applying condition (b) to the point p and any point on the CTC, $\partial I^-(\text{CTC})$ must have compact slices on the globally hyperbolic part of the manifold. Thus the generalized entropy should increase towards the future on $\partial I^-(\text{CTC})$ on a compact slice.

But by asymptotic flatness, this compact horizon must be contracting at early times, violating the GSL for the same reason that baby universes did in section 4.1. Consequently, assuming the GSL and condition b (which is the other half of global hyperbolicity), it is impossible for any experimenter sitting at a point p in an asymptotically flat universe to arrange for a CTC to form. Similarly, the time-reverse of the above argument using $\overline{\text{GSL}}$ shows that if there are currently existing time machines it is impossible for them to be destroyed.

5 Does It Still Work for Quantum Gravity?

The above results have been proven on the assumption that spacetime can be approximated by a smooth, globally-hyperbolic, Lorentzian manifold subject to a small quantum perturbation, such that the resulting spacetime satisfies the GSL exactly. The question is whether we expect the result to hold even if some of these assumptions are relaxed. In particular, the following questions arise:

What about thermodynamic fluctuations which cause the entropy to temporarily decrease, thus making the GSL not exact? Do quantum fluctuations in the metric make the GSL ill-defined in the quantum gravity regime? And what about global hyperbolicity and the other assumptions involved in the no-go results of section 4?

5.1 Entropy fluctuations

All thermodynamic systems have fluctuations, since it is always possible that the degrees of freedom in a complex system will, by chance, temporarily enter an unlikely configuration. When the entropy is defined as the Boltzmann entropy $S = \ln N$, where N is the number of microstates in a macrostate, this can lead to a temporary decrease in the entropy. However, in the approach to the second law which is used here, the entropy of quantum fields is defined in Eq. (2) using the Gibbs entropy $S = -\mathrm{tr}(\rho \ln \rho)$. As shown in section 2.1, this entropy can be proven to be exactly nondecreasing. In the case of the GSL one also has the area term. I have argued elsewhere [28] that one can take a similar interpretation of the generalized entropy if one defines the GSL using the expectation value $\langle A \rangle$, as suggested in Ref. [55]. The GSL as defined in this way need not have any downward fluctuations, which is convenient for proofs of the GSL.

Nevertheless, changing the definition of the entropy cannot change the underlying physics, and the entropy fluctuations are still present and physically important. In the Gibbs interpretation, these entropy fluctuations appear when one invokes the probability interpretation of the density matrix ρ. For example, consider the spin of an electron which has two states,up and down. If the system is in uniformly mixed density matrix with diagonal (1/2, 1/2), the entropy is In 2. But this density matrix only represents our ignorance; the electron may well actually be in the up state. And it is easy to show that the maximum entropy associated with any pure state is 0. The following superficially valid syllogism is therefore fallacious:

1. If the electron is in the up state, $S = 0$.
2. The entropy of the electron is $S = \ln 2$,
3. Therefore the electron is not in the up state.

The correct conclusion is that the electron *might* not be in the up state, a very different statement.

Similarly, if the generalized entropy decreases somewhere on a null surface N, the proper conclusion to draw is not that N is not a horizon, but that N *might not* be a horizon. Let us take as a specific example the no-traversable-wormholes result from section 4.1. A sufficiently adventurous spacefarer might not be deterred from attempted to crossing a wormhole simply because it is uncertain whether or not he will

make it. Suppose then that our intrepid hero lives in an asymptotically flat universe with state Ψ, and then jumps into the wormhole even though he only has a probability $1 > p > 0$ of successfully reaching the other side. Assuming he *does* reach the other side, it is then appropriate, at least on a forward-going basis, to project the state of the universe onto a new state Ψ' in which the wormhole jump certainly occurs, by using the projection operator P onto the fact of the wormhole traversal:

$$|\Psi'\rangle = \frac{P}{\sqrt{p}}|\Psi\rangle = \sqrt{p}\,|\Psi + \sqrt{1-p}\,|\chi\rangle, \tag{35}$$

where χ is some orthogonal state.

Here it is necessary to be careful. As usual in quantum mechanical measurement, the state Ψ' will not be a good description of the state of the universe prior to the time of measurement. In fact, since the additional branch χ of the superposition is defined using a future boundary condition, by the arguments in section 2.1, one expects the coarse-grained ordinary entropy of χ to be decreasing with time prior to the act of measurement. However, it is not necessary to insist on Ψ' being the true state of the universe, or on wavefunction collapse being the correct interpretation of quantum mechanical measurements. It is only required that Ψ' be a well-defined state in the theory, to which the fine-grained GSL must therefore apply. One expects that Ψ' will be asymptotically flat since this boundary condition should not be affected by anything which goes on in the interior of spacetime. Then Ψ' is an asymptotically flat spacetime in which there exists a traversable wormhole with probability 1, contradicting the result in section 4.1. Similar arguments apply to the other trapped-surface no-go results in section 4.[19]

It therefore follows from the GSL that the probability of forming a baby universe, a traversable wormhole, or restarting inflation in asymptotically flat or AdS spacetime, is *exactly* zero. This is a little surprising because one might have thought that these things could occur through quantum tunneling [48, 46]. However, it is in accordance with the observation in Ref. [47] that one cannot restart inflation in AdS spacetimes in the context of AdS/CFT.

5.2 Quantum geometries

Since singularities are regions where quantum gravity effects might become important, one critical question is whether the GSL (as defined in section 2.2) continues to remain well-defined and true in the quantum gravity regime. If not, then the quantum singularity theorem might break down just when it is needed, although many of the other applications of the Penrose theorem would continue to be useful. Consider the following hierarchy of increasingly "quantum" treatments of spacetime:

1. Weak semiclassical perturbations. This regime is simply quantum field theory in curved spacetime, plus infinitesimal corrections due to the gravitational deformation from matter. (This regime justifies the 'semiclassical expansion' used in Theorems 1, 3 &4 of section 3.)

2. Strong semiclassical effects. In this regime one permits the quantum fields to have large gravitational effects, but neglects any quantum fluctuations in the metric, so that spacetime is still described by a smooth Lorentzian manifold. This regime can be justified in some cases when there are a large number N of matter fields and one takes $\hbar \to 0$ while holding $N\hbar$ fixed. This suppresses graviton loops relative to matter loops.
3. Quantum Lorentzian manifold. In this regime one takes into account the fluctuations in the metric, treating the metric as a quantum field living on a fixed, continuous topological space. Some of the geometrical quantities now fail to commute. This regime includes spacetimes which can be described by perturbative quantum gravity. It is even possible, if the asymptotic safety scenario is true, that this regime might encompass a complete theory of quantum gravity [57].
4. Something New. Examples include causal sets [58], the discrete spacetimes of loop quantum gravity [59] or matrix theory [60], etc. Here one cannot say anything nonspeculative except on the assumption of a specific model.

Which of these regimes is physically relevant for the no-go results?

Not all of the results in this article require probing the quantum gravity regime. For example, the absence of traversable wormholes (section 4.1) or warp drives (section 4.3) is interesting even perturbatively, in the weak semiclassical regime (#1). That is because there exist classical solutions that are right on the edge of violating these results, so that a violation could be seen even perturbatively. (For warp drives, consider pertrubations to the vacuum solution. For traversable wormholes, perturb the eternal black hole.)[20]

But the usefulness of the no-go results concerning singularity theorems and baby universes (sections 4.1–4.2) depends on their applicability to the region near the singularity. Consider a spacetime whose classical evolution has a singularity, and suppose that some resolution of the singularity *were* possible. Necessarily, any resolution of the singularity would have to involve nonperturbative effects, and therefore the weak semiclassical regime (#1) will be insufficient to resolve the singularity.[21]

However, there is still hope for a useful result. The key thing to notice is that these results involve two distinct locations on the same null-surface N, separated by some null interval $\Delta\lambda$. The region X near the quantum trapped surface may be weakly coupled, even while the classically singular region Y is strongly coupled. Near X, it can be shown that the generalized entropy is decreasing somewhere on the null surface. Then the GSL implies that N cannot be a causal horizon. Near Y, we use the fact that N is not a horizon to show that there can be no infinite worldline W in Y.

For example, in the case of the Big Bang result, the region Y is the region near[22] the classical would-be initial singularity when the universe was very hot and dense, while X may be taken to be the present-day cosmology, which is very well-described by classical general relativity. The analysis of region X by itself can therefore be carried out entirely in the weak semiclassical regime (#1). Thus, the applications of Theorems 1 and 2 to region X are insensitive to quantum gravity effects. X is also the only location at which we used a formula for the generalized entropy S_{gen}, so quantum gravity corrections to the formula for S_{gen} are irrelevant to the validity of the result.

This leaves the analysis of the strongly coupled region *Y,* which might be in any of regimes #2 – 4. In order for the GSL to be well defined, it is necessary that the concept of a causal horizon still exist. The notion is clearly defined in the strong semiclassical regime (#2) due to the existence of a Lorentzian spacetime. For a quantum Lorentzian manifold (#3), the causal structure becomes fuzzy and thus one might worry about whether the causal horizon is defined. Let us assume the following correspondence principles: A) that the requirement in general relativity that coordinates be smooth is merely for technical convenience, and that therefore one can consistently formulate general relativity to be covariant under the choice of arbitrary continuous coordinates, not just diffeomorphisms, B) that for any way of consistently gauge-fixing classical general relativity, there is a corresponding way to gauge-fix a quantum Lorentzian manifold, without introducing an anomaly into the true diffeomorphism-invariance of the theory.

Let there be some locus of points L defined by some generally covariant prescription. Then $\partial I^-(L)$ is a continuous (but not necessarily smooth) surface of codimension 1. By assumption A above it is consistent to gauge-fix general relativity using a coordinate system in which one coordinate x satisfies $x = 0$ at $\partial I^-(L)$. In this coordinate system the horizon location is in a well-defined, fixed position in space, and therefore does not fluctuate in its position. If L is taken to be a future-infinite worldline defined by any coordinate invariant prescription, this shows that the notion of a future horizon is well-defined. The GSL can then be defined to require that any such future causal horizon have nondecreasing entropy, at least in the semiclassical region X where we know how to define S_{gen}.

Even if the spacetime geometry is described by some new discrete structure (#4), it still seems reasonable to believe that the notion of a future horizon may be well-defined, *if* this discrete structure has a fundamental notion of causality built into it. One can think of the location of a causal horizon as being defined by the way in which it divides spacetime points into exterior and interior regions. The exterior of a future horizon can be defined as $J^-(W_{\text{fut}})$, the causal past of a future-infinite worldline. A worldline W can be defined as a chain of points in causal sequence. The only part of the definition which depends on anything other than a causal structure, is the requirement that W_{fut} be infinite. But this can naturally be defined in a discrete geometry by requiring W to consist of an infinite number of points.

So if the quantum geometry of the universe is a discrete causal set or anything richer, the notion of a future horizon should be well-defined, and it should be possible to ask whether the GSL is true. Of course, it might turn out to be false. For example, if quantum gravity violates Lorentz invariance, then the arguments of Ref. [61] suggest that the GSL will be invalid.

However, it seems more elegant for the GSL to be true in quantum gravity. This would explain the success of horizon thermodynamics in semiclassical general relativity. Not only that, but by the results in sections 4.3 and 4.4, it would also ensure that the theory has positive energies and good causality.

Global Hyperbolicity. About half of the results in section 4 assume global hyperbolicity, either directly or through the use of Theorems 2 or 4. This includes

the generalization of the Penrose singularity theorem (sections 4.1 and 4.2), some of the discussion about the thermodynamic beginning of the universe (4.2), and the prohibition of warp drives, and negative mass objects (4.3). On the other hand, the prohibitions on viable baby universes, traversable wormholes, and restarting inflation (4.1) use only the GSL, while the no-time machines result (4.4) uses a weakened form of global hyperbolicity.

If the theory of quantum gravity is fully predictive, one expects some analogue of global hyperbolicity to be true, but it may not have quite the same implications as in general relativity. Just because topology change is forbidden for continuous globally hyperbolic manifolds [45], does not necessarily mean it could not occur in discrete spacetimes. Thus it is necessary to examine whether one expects theorems 2 and 4 to continue to hold. Theorem 2 only depends on global hyperbolicity insofar as this is necessary to identify causal subsystems; it is therefore likely to hold in any theory with causality.

Theorem 4 depends on global hyperbolicity in a more subtle way. The basic causality assumption underlying Theorem 4 is that an outward moving causal surface on a noncompact spatial slice cannot come to an end without encountering a boundary of the spacetime. This might happen in two different ways: (a) A noncompact space could become compact as a result of time evolution, or (b) A noncompact space could split into two regions, one of them compact, and the other noncompact. It seems unlikely that quantum topology change could permit (a), since it would require an "infinitely large" tunneling event. Scenario (b) is the disconnected baby universe scenario, which is forbidden by the GSL *without* using global hyperbolicity.[23] Accepting these arguments against (a) and (b), it is not unlikely that an analogue of Theorem 4 may well apply in full quantum gravity.

Thus there is a reasonable possibility that the Penrose singularity theorem can be proven even in the context of full quantum gravity. This would go against the conventional wisdom that the singularities are an symptom of the incompleteness of the classical theory, and are resolved quantum mechanically. However, it should be pointed out that just because there are singularities in the sense that spacetime comes to an end in some places, does not mean that there are any physical quantities which become infinite at the singularity. A discrete geometry might still resolve the singularity in the latter sense by cutting off the spacetime at distances shorter than the Planck scale.

6 Conclusion

It has been shown above that, under the assumption that spacetime is a globally hyperbolic Lorentzian manifold, the fine-grained GSL requires black holes and infinite FRW universes to have singularities, and places severe constraints on baby universes and any cosmology prior to the Big Bang. It additionally prevents asymptotically flat spacetimes from having negative ADM masses, warp drives or traversable wormholes, or developing time machines or inflating regions. In all of these cases, theorems of classical general

relativity have been extended to semiclassical settings by using the GSL as a premise instead of the null energy condition. The notion of a "trapped surface" still persists in this quantum setting, and ensures that these qualitative features of semiclassical gravity are the same as those of classical general relativity. (It should be reiterated, however, that the GSL has only been proven in limited regimes [28], and that there might be other reasonable ways to formulate the GSL besides the one given in section 2.2.)

There are also some—necessarily speculative—indications that these results might hold in a full theory of quantum gravity. Although the semiclassical approximation was used to derive some of the no-go results, it was only used in nearly classical regions, either a large distance or a long time away from high curvature quantum gravity regions. Other than the GSL itself, the only assumptions made about the high curvature region were that spacetime continues to have some of the same primitive properties as a Lorentzian spacetime: a notion of causality used to define horizons, a notion of predictivity analogous to global hyperbolicity, distinctions between finite and infinite lengths, and compact and noncompact regions. Given the successes of horizon thermodynamics, it is natural to suppose that the GSL holds even at the level of quantum gravity, and thus that not all singularities are resolved in quantum gravity.

The statistical mechanical argument for a beginning in time, based on the fact that entropy decreases when going to the past, was also generalized to an argument from the coarse-grained GSL. Together with the singularity theorem, this leads to a prima facie argument that time had a true beginning at the Big Bang some 13.7 billion years ago. In section 4.2, in order to make a plausible GSL-satisfying cosmology with an infinite past, it was necessary to postulate both that the cosmos is spatially finite, and that the arrow of time was reversed before some time t_0. This kind of bounce evades both the singularity and thermodynamic arrow constraints, but still has in some sense a thermodynamic 'beginning' in time at the moment of lowest entropy. That is, both the past and the future would be explained in terms of the low entropy state at t_0, while the state at t_0 would itself have no explanation in terms of anything to the future or the past. (Thus the moment t_0 would seem to raise the same sorts of philosophical questions that any other sort of beginning in time would.)

The fact that the no-go results forbid various processes with probability 0 is interesting because it goes against the usual experience in quantum field theory that anything not forbidden by kinematics or conservation laws must occur with some nonzero probability. This suggests that there may be a formulation of quantum geometry based on horizon thermodynamics in which these constraints seem more natural.

The notion of a quantum trapped surface from Theorem 4 may be a clue here. If we think that horizon thermodynamics works because of the statistical mechanics of the quantum gravity degrees of freedom near or on the horizon, what should we make of the fact that on certain surfaces, the entropy does decrease? It is as though each null surface must either choose to be a causal horizon and behave in certain respects like a closed system, or else violate the second law and be punished for it by coming to an end in a finite time. Can this basic dichotomy be explained somehow from the perspective of the microscopic horizon degrees of freedom?

Acknowledgments

This work was supported by NSF grants PHY-0601800 and PHY-0903572, the Maryland Center for Fundamental Physics, the Perimeter Institute for Theoretical Physics, and the Simons Foundation. I am grateful for conversations with Ted Jacobson, William Donnelly, Sudipta Sarkar, Latham Boyle, Rob Myers, Don Marolf, Robert Brandenberger, Netta Engelhardt, Ed Witten, and for detailed comments from anonymous referees.

Appendix

This appendix will prove a theorem used in section 2.1 to prove that entropy increases in the context of ordinary thermodynamics. There, the approach was to model our uncertainty about time evolution using a mixture of possible unitary operators, acting on any separable Hilbert Space.

Let dU be a probability distribution over the space of possible unitary operators, such that the total probability is 1:

$$\int_U dU = 1 \tag{36}$$

Then the operation to be performed on the state ρ is

$$\rho \to T(\rho) \equiv \int_U U \rho U^\dagger dU. \tag{37}$$

T is a linear map from the space of all density matrices to itself. There are some consequences of the fact that the map is just a sum over unitary operators: a) By conservation of probability, T must preserve the trace of ρ. b) Since probabilities cannot be negative, T must map states with nonnegative eigenvalues to other states with nonnegative eigenvalues. Furthermore, c) Since every U preserves the identity state I, $T(I) = I$.

Such maps cannot decrease the entropy, a fact which has been derived from the Uhlmann theory of mixing [20], by way of an even stronger statement about probability eigenvalues: namely that there is no way of transferring probability from a smaller eigenvalue to a larger one.

More precisely, for all natural numbers i, the sum of the ith largest eigenvalues of the final density matrix must be no greater than the ith largest eigenvalues of the initial density matrix. In other words, if we consider the spectrum of probability eigenvalues of ρ, a probability eigenvalue can only increase in value if it does so at the expense of eigenvalues with greater probability. This stronger statement is more powerful than simply asserting that entropy increases, because it yields a separate statement for each number i.

One might worry about how to interpret "the sum of the ith largest eigenvectors" if ρ is degenerate. There is then an ambiguity as to a) which basis to use to count

eigenvectors, and b) how to order the eigenvalues seeing as some of them are equal. However, because by definition all of the degenerate eigenvalues have the same value, it makes no difference to the sum how we partition the degenerate states so long as we make some choice.

A proof of this result follows:

Definitions:

ρ is the density matrix,
$T()$ is the trace and identity preserving, positive linear map,
I is the identity matrix,
$M = \rho - p_i I$, where p_i is the ith largest eigenvalue,
P projects onto the ith largest eigenvalues of ρ, and
Q projects onto the ith largest eigenvalues of $T(\rho)$.

Theorem 5: The theorem to be proven can now be stated as follows:

$$\mathrm{tr}\left(QT\left(\rho\right)Q\right) \leq \mathrm{tr}\left(P\rho P\right) \tag{38}$$

Proof: Since P commutes with ρ and thus M, we may write

$$M = PMP + \left(1-P\right)M\left(1-P\right) = A - B \tag{39}$$

where A and B are manifestly positive. Now by positivity,

$$\mathrm{tr}\left(QT\left(A\right)Q\right) \leq \mathrm{tr}\left(T\left(A\right)\right) = \mathrm{tr}\left(A\right) \tag{40}$$

since a partial trace of a positive matrix cannot give more than the full trace, while

$$\mathrm{tr}\left(QT\left(B\right)Q\right) \geq 0 \tag{41}$$

because each operation preserves positivity. Therefore by linearity,

$$\mathrm{tr}\left(QT\left(M\right)Q\right) \leq \mathrm{tr}\left(A\right) = \mathrm{tr}\left(PMP\right) \tag{42}$$

which shows that the sum of the i larger eigenvalues can only decrease. But M and ρ only differ by $I = T(I)$ so the result holds for ρ as well. Q.E.D.

Corollary 5.1: Any quantity expressible as $\mathrm{tr}(f(\rho))$, where f is any convex function, is nondecreasing. This is implied by the fact that the probability eigenvalues can only evolve towards equalization [20].

This stronger statement about probability eigenvalues is equivalent to saying that any convex function of the probability eigenvalues is nondecreasing. Pick the convex function:

$$f\left(p\right) = 0 \quad p \leq p_i \tag{43}$$

$$= p_i - p \quad p \geq p_i \tag{44}$$

This function must satisfy

$$\operatorname{tr} f\left(T\left(\rho\right)\right) \geq \operatorname{tr} f\left(\rho\right) \tag{45}$$

which implies that the probability eigenvalues can only equalize.

Notes

1 A set is achronal if no two points are connected by timelike curves.
2 They could not prove any singularity theorems, because these typically require the averaged null energy condition to hold on a semi-infinite null ray with one endpoint, a condition which can be violated by quantum fields.
3 While quantum gravity effects are expected to be important near singularities, other applications of the Penrose theorem do not require them (cf. section 5.2).
4 Or more generally, CPT symmetry.
5 It is also necessary not to make any restriction on the final state of the universe. If there were a low entropy assumption made for both the initial and the final state, it would not be correct to calculate $\tilde{\rho}$ from the initial condition alone, since that would ignore additional relevant information. This assumption is implicit in the argument for the OSL given above.
6 Here we are talking about the *ordinary* entropy of the interior of the black hole, not the *generalized* entropy of its horizon (which is the subject of the next section). The latter is bounded at any finite energy; the former might not be, depending on one's views about entropy bounds and black hole information loss. My claims in this article do not require taking a stand on this controversy, since they will be based on the generalized entropy rather than the ordinary entropy.
8 Technically the laws of physics are invariant under CPT, not T by itself. This does not affect the argument because the generalized entropy is invariant under C and P. But if the laws of quantum gravity were to violate CPT, the GSL and its time-reverse might be independent of each other.
9 This is distinct from the semiclassical approximation involving a large number N of species, in which $N\hbar$ is held fixed as one takes the $\hbar \to 0$ limit, which will be discussed in section 5.2.
10 A typical null surface will develop cusps where its generators enter or leave the surface, and at these points the surface will not be smooth. On a smooth spacetime these nonsmooth parts of the null surface are usually of lower dimension, so this assumption is reasonable if the point g_0 is generic.
11 Both $\sum$ and $\sum'$ are assumed to be approximately constant over the length scale set by the proper distance f—otherwise one could satisfy Theorem 1 simply by evolving forwards in time on only one of the two null surfaces N or M!
12 Traditionally, global hyperbolicity requires also that the spacetime satisfy strong causality, but this apparently stronger form of global hyperbolicity was recently proven from the definition given here [43].

13 Technically, this definition excludes baby universes which are eventually end in a Big Crunch without spawning any new universes themselves. However, the no-go theorem might possibly be extended to such cases by arguing like this: As long as the baby universe expands for a long time without recollapse, there exist approximate W_{fut}'s in the form of worldlines which exist for a very long time without collapse. Since the GSL holds for infinite W_{fut}, by continuity there ought to be some sense in which the GSL is very close to true for very long but finite worldlines.

14 J^- has been used instead of I^- in order to make this region closed.

15 There are also singularity theorems which apply to spatially finite universes [6], but these theorems use the strong energy condition. This condition can be violated even by classical, minimally-coupled scalar fields (and was violated in the early universe, if inflationary cosmology is true). There is little reason to believe that either these theorems or some quantum analogue apply to the early universe.

16 By the generic condition used in the Corollary 2.1, these bubbles are bounded by future-trapped surfaces. By Theorem 3, they must eventually contract to nothing. If the FRW cosmology has no final singularity, this must happen simply by the lightrays crossing each other. Therefore, there exists a complete spatial slice Σ in the FRW cosmology with the property that the entire Σ lies to the future of any past infinite worldline W_{past}.

17 In a bouncing scenario, it is not logically necessary that the moment of lowest entropy was the same as the moment of smallest size, but it seems natural to make this identification.

18 If you think that a circular worldline should not count as "infinite" in the relevant sense, simply consider a slightly wiggly line near the CTC which never exactly intersects itself.

19 In contrast, entropy fluctuations are relevant when applying the GSL to inflationary cosmology. In an inflationary scenario, there is a scalar field Φ with some potential $V(\Phi)$, which gives rise to vacuum energy and a de Sitter-like exponential expansion. As this scalar field rolls down the potential, inflation comes to a halt. The argument of eternal inflation is that quantum fluctuations sometimes push the field back up the potential and thus increase rather than decrease the vacuum energy [56]. Since an upward fluctuation results in a decrease in the horizon area, naively it would seem that the GSL forbids this process as well. If so, the vacuum energy would be nonincreasing as time passes. And if it cannot increase, one would generically expect it to decrease, and eventually exit inflation everywhere. Assessing the validity of this argument requires a careful consideration of vacuum fluctuations. The argument above that entropy fluctuations do not matter applies to asymptotically flat spacetimes. De Sitter space is different for two reasons: a) In de Sitter space there is a maximum value of the generalized entropy; so the necessary downward fluctuation ΔS of the horizon entropy is finite rather than infinite. b) In order to have eternal inflation, it is not necessary that any pre-selected region of spacetime remain inflating, but only that there exists some region that continues to do so. If, over a given time interval, a Hubble volume increases in volume by a factor of N, it is only necessary to have a probability of about $1/N$ that inflation continue in each Hubble volume in order to keep inflation going somewhere. If $\ln N > \Delta S$, then one would expect the necessary entropy fluctuation to occur in one of the N regions. It would be interesting to check whether this condition places significant constraints on eternal inflation scenarios.

20 Of course, if these results continue to hold even in even more quantum spacetimes (regimes #2-4), for the reasons suggested below, so much the better.

21 As a general mathematical fact, when singularities in some function are resolved, the resolution tends to be nonperturbative. At any finite order in perturbation theory the singularity typically gets more divergent rather than less divergent. A simple example: if you Taylor expand the function $f(x) = 1/(x^2 + a^2)$ in a around $a = 0$, each term in the Taylor series is progressively *more* divergent with respect to x, but at finite positive values of a, there is no singularity. I would like to thank Ed Witten for pointing out this issue.

22 And before, if the singularity is resolved.

23 Additionally, (b) raises potential problems with causality. If one runs the process in time-reverse, one finds that two completely unrelated regions of spacetime spontaneously join together. This seems to be an extreme violation of locality. But see Ref. [62] for a possible way around this argument.

References

[1] B.S. DeWitt, "Quantum Theory of Gravity. I. The Canonical Theory", Phys. Rev. **160**, 1113 (1967); C.W. Misner, K.S. Thorne, J.A. Wheeler, *Gravitation*, W.H. Freeman and Company 1973.

[2] G. Veneziano, "String Cosmology: The Pre-Big Bang Scenario", arXiv:hep-th/0002094v1 (2000); J. Khoury, B.A. Ovrut, N. Seiberg, P.J. Steinhardt, N. Turok, "From Big Crunch to Big Bang", Phys. Rev. D **65,** 086007 (2002), arXiv:hep-th/0108187; M. Gasperini, G. Veneziano, "The Pre-Big Bang Scenario in String Cosmology", Phys. Rept. **373,** 1 (2003), arXiv:hep-th/0207130.

[3] M. Bojowald, "Absence of a Singularity in Loop Quantum Cosmology", Phys. Rev. Lett. **86,** 5227 (2001), arXiv:gr-qc/0102069; A. Ashtekar, "Singularity resolution in loop quantum cosmology: A brief overview"; J. Phys. Conf. Ser. **189,** 012003 (2009), arXiv:0812.4703v1.

[4] R.H. Brandenberger, "Cosmology of the Very Early Universe", arXiv:1003.1745v1, and citations therein.

[5] R.D. Sorkin, "On the entropy of the vacuum outside a horizon", talk given at the proceedings of the GR10 conference in Padova, 1983; V. Frolov, I. Novikov, "Dynamical origin of the entropy of a black hole", Phys. Rev. D **48,** 4545 (1993), arXiv:gr-qc/9309001; L. Susskind, J. Uglum, "Black hole entropy in canonical quantum gravity and superstring theory", Phys. Rev. D **50,** 2700 (1994), arXiv:hep-th/9401070; T. Jacobson, "Black hole entropy and induced gravity" (1994), arXiv:gr-qc/9404039; A. O. Barvinsky, V. P. Frolov, A. I. Zelnikov, "Wavefunction of a black hole and the dynamical origin of entropy", Phys. Rev. D **51,** 1741 (1995), arXiv:gr-qc/9404036; V. P. Frolov, D. V. Fursaev, A. I. Zelnikov, "Statistical origin of black hole entropy in induced gravity", Nucl. Phys. B **486,** 339 (1997), arXiv:hep-th/9607104.

[6] S.W. Hawking, G.F.R. Ellis, *The large-scale structure of space-time*, Cambridge University Press 1973. See also R.M. Wald, *General Relativity*, University of Chicago Press 1984.

[7] M.S. Morris, K.S. Thorne, U. Yurtsever, "Wormholes, time machines, and the weak energy condition", Phys. Rev. Lett. **61,** 1446 (1988); J.L. Friedman, K. Schleich, D.M. Witt, "Topological censorship", Phys. Rev. Lett. **71,** 1486 (1993), arXiv:gr-qc/9305017v2;

[8] E. Farhi, A.H. Guth, "An obstacle to creating a universe in the laboratory" Phys. Lett. B **183**, 149 (1987).

[9] F.J. Tipler, "Causality Violation in Asymptotically Flat Space-Times", Phys. Rev. Lett. **37**, 879 (1976); S.W. Hawking, "Chronology protection conjecture", Phys. Rev. D **46,** 603 (1992).

[10] K.D. Olum, "Superluminal travel requires negative energies", Phys. Rev. Lett. **81**, 3567 (1998), arXiv:gr-qc/9805003v2.

[11] M. Visser, B.A. Bassett, S. Liberati, "Superluminal censorship", Nucl. Phys. Proc. Suppl. **88**, 267 (2000), arXiv:gr-qc/9810026v2. 20.

[12] E. Woolgar, "The Positivity of Energy for Asymptotically Anti-de Sitter Space-times", Class. Quant. Grav. 11, 1881-1900 (1994), arXiv:gr-qc/9404019.

[13] S. Gao, R.M. Wald, "Theorems on gravitational time delay and related issues", Class. Quant. Grav. **17**, 4999 (2000), arXiv:gr-qc/0007021.

[14] R. Penrose, R.D. Sorkin, E. Woolgar, "A positive mass theorem based on the focusing and retardation of null geodesics", arXiv:gr-qc/9301015v2.

[15] L.S. Brown, G.J. Maclay, "Vacuum stress between conducting plates: an image solution", Phys. Rev. **184,** 1272 (1969); H. Epstein, V. Glaser, A. Jaffe, "Non-positivity of the energy density in quantized field theories", Nuovo Cim. **36,** 1016 (1965); S.A. Fulling, P.C.W. Davies, "Radiation from a moving mirror in two dimensional space-time: conformal anomaly", Proc. R. Soc. Lond. A **348,** 393 (1976); P.C.W. Davies, S.A. Fulling, "Radiation from moving mirrors and from black holes", Proc. R. Soc. Lond. A **356,** 237 (1977). For S.L. Braunstein's discovery that squeezed states of electromagnetic fields can have negative energy density, see M.S. Morris, K.S. Thorne, "Wormholes in spacetime and their use for interstellar travel: A tool for teaching general relativity", Am. J. Phys **56,** 5 (1988).

[16] J.B. Bekenstein, "Nonsingular general-relativistic cosmologies", Phys. Rev. D **11**, 2072 (1975); C. Barcelo, M. Visser, "Twilight for the energy conditions?", Int. J. Mod. Phys. D **11,** 1553 (2002), arXiv:gr-qc/0205066v1.

[17] N. Graham, K.D. Olum, "Achronal averaged null energy condition", Phys. Rev. D **76**, 064001 (2007), arXiv:0705.3193v2.

[18] A.C. Wall, "Proving the Achronal Averaged Null Energy Condition from the Generalized Second Law", Phys. Rev. D **81,** 024038 (2010), arXiv:0910.5751v2.

[19] M. Delbrück, G. Molière, Abhand. Preuss. Akad. Wiss. Phys. Math. **1**, 25 (1936), as cited by L.W. Bruch, H. Falk, "Gibbs Inequality in Quantum Statistical Mechanics", Phys. Rev. A **2,** 1598 (1970). See also H. Falk, E. Adler, "Quantum Counterpart of Gibbs Theorem on Entropy and Statistical Dependence", Phys. Rev. **168,** 185 (1968).

[20] Uhlmann, A., Wiss. Z. Karl-Marx-Univ. Leipzig **20,** 633 (1971), id. **21,** 427 (1972), id. **22,** 139 (1973), as cited by A. Wehrl, "General Properties of Entropy", Rev. Mod. Phys. **50,** 221 (1978); A. Wehrl, "How chaotic is a state of a quantum system?" Rep. Math. Phys. **6,** 15 (1974).

[21] S.M. Carroll, J. Chen, "Spontaneous Inflation and the Origin of the Arrow of Time" (2004), arXiv:hep-th/0410270v1.

[22] T.M. Fiola, J. Preskill, A. Strominger, S.P. Trivedi, "Black hole thermodynamics and information loss in two dimensions", Phys. Rev. D **50**, 3987 (1994), arXiv:hep-th/9403137.

[23] J. A. Wheeler, 1974. "Beyond the end of time", in M. Rees, R. Ruffini, J.A. Wheeler, *Black Holes, Gravitational Waves and Cosmology: An introduction to current research.*, Gordon and Breach 1974.

[24] V. P. Frolov, M. A. Markov, M. A. Mukhanov. "Through a black hole into a new universe?", Phys. Lett. B **216**, 272 (1989).

[25] L. Smolin, "Did the Universe Evolve?" Class. Quantum Grav. **9,** 173 (1992), "The fate of black hole singularities and the parameters of the standard models of particle physics and cosmology" (1994), arXiv:gr-qc/9404011, "The status of cosmological natural selection" (2006), arXiv:hep-th/0612185v1.

[26] J. Khoury, B.A. Ovrut, P.J. Steinhardt, N. Turok, "Ekpyrotic universe: Colliding branes and the origin of the hot big bang", Phys. Rev. D **64**, 123522 (2001), arXiv:hep-th/0103239v3.

[27] W. G. Unruh, R. M. Wald, "Acceleration radiation and generalized second law of thermodynamics," Phys. Rev. D **25,** 942 (1982); "Entropy bounds, acceleration radiation, and the generalized second law," Phys. Rev. D **27,** 2271 (1983); M. J. Radzikowski, W. G. Unruh, "Correction to 'Acceleration radiation and the generalized second law of thermodynamics'," Phys. Rev. D **37,** 3059 (1988); D. Marolf, R. Sorkin "On the status of highly entropic objects" Phys. Rev. D **69,** 024014 (2004), arXiv:hep-th/0309218; D. Marolf, R. Roiban, "Note on bound states and the Bekenstein bound", JHEP 08, 033 (2004), arXiv:hep-th/0406037; D.N. Page, "Hawking radiation and black hole thermodynamics", New J. Phys. **7,** 203 (2005), arXiv:hep-th/0409024.

[28] A.C. Wall "Ten Proofs of the Generalized Second Law", JHEP 0906**,** 021 (2009), arXiv:0901.3865v2, and references.

[29] A.C. Wall "A proof of the generalized second law for rapidly evolving Rindler horizons", Phys.Rev. D **82,** 124019 (2010), arXiv:1007.1493; A.C. Wall, "A proof of the generalized second law for rapidly changing fields and arbitrary horizon slices", Phys. Rev. D **85,** 104049 (2012), arXiv:1105.3445v2.

[30] R.M. Wald, "Black hole entropy is Noether charge", Phys. Rev. D **48,** 3427 (1993), arXiv:gr-qc/9307038; V. Iyer, R.M. Wald, "Some Properties of Noether Charge and a Proposal for Dynamical Black Hole Entropy", Phys.Rev. D **50,** 846 (1994).

[31] S.W. Hawking, "Particle creation by black holes", Commun. Math. Phys. **43**, 199 (1975).

[32] D.V. Fursaev, S.N. Solodukhin, "On one-loop renormalization of blackhole entropy", Phys. Lett. B **365,** 51 (1996), arXiv:hep-th/9412020; J.-G. Demers, R. Lafrance, R.C. Myers, "Black hole entropy and renormalization" (1995), arXiv:gr-qc/9507042; S.N. Solodukhin, "One-loop renormalization of black hole entropy due to non-minimally coupled matter", Phys. Rev. D **52,** 7046 (1995), arXiv:hep-th/9504022; S.P. de Alwis, N. Ohta, "Thermodynamics of quantum fields in black hole backgrounds," Phys. Rev. D **52,** 3529 (1995), arXiv:hep-th/9504033; E. Winstanley, "Renormalized black hole entropy in anti-de Sitter space via the 'brick wall' method", Phys. Rev. D **63,** 084013 (2001), arXiv:hep-th/0011176. For some two-dimensional results, see V.P. Frolov, D.V. Fursaev, A.I. Zelnikov, "Black hole entropy: thermodynamics, statistical-mechanics and subtraction procedure", Phys. Lett. B **382,** 220 (1996), arXiv:hep-th/9603175; eid. "Black hole entropy: off-shell vs on-shell", Phys. Rev. D **54,** 2711 (1996), arXiv:hep-th/9512184v2. For a seeming discrepency for scalar fields in odd dimensions, see Kim, Kim, Soh, Yee, "Renormalized thermodynamic entropy of black holes in higher dimensions", Phys. Rev. D 55, 2159 (1997), arXiv:gr-qc/9608015v3.

[33] D. Kabat, "Black hole entropy and entropy of entanglement" Nucl. Phys. B **453,** 281 (1995), arXiv:hep-th/9503016; F. Larsen, F. Wilczek, "Renormalization of black hole

entropy and of the gravitational coupling constant", Nucl. Phys. B **458,** 249 (1996), arXiv:hep-th/9506066; D.V. Fursaev, G. Miele, "Cones, Spins and Heat Kernels", Nucl. Phys. B **484,** 697 (1997), arXiv:hep-th/9605153; W. Donnelly, A.C. Wall, "Do gauge fields really contribute negatively to black hole entropy?", Phys. Rev. D **86,** 064042 (2012), arXiv:1206.5831.

[34] T.A. Jacobson, R. Parentani, "Horizon entropy", Found. Phys. **33,** 323 (2003), arXiv:gr-qc/0302099.

[35] See Appendix B of Ref. [22], and discussion in Ref. [28].

[36] R.D. Sorkin, "Ten theses on black hole entropy", Stud. Hist. Philos. Mod. Phys. **36**, 291 (2005), arXiv:hep-th/0504037.

[37] Ted Jacobson, Robert C. Myers, "Entropy of Lovelock Black Holes", Phys. Rev. Lett. **70,** 3684 (1993); T. Liko, Phys. Rev. D **77,** 064004 (2008), arXiv:0705.1518; S. Sarkar, A.C. Wall, tentative title "Second Law Violations in Lovelock Gravity", forthcoming.

[38] T. Jacobson, G. Kang, R.C. Myers, "Increase of black hole entropy in higher curvature gravity", Phys. Rev. D **52**, 3518 (1995), arXiv:gr-qc/9503020.

[39] C.P. Burgess, "Quantum Gravity in Everyday Life: General Relativity as an Effective Field Theory", Living. Rev. Rel. **7,** 5 (2004), arXiv:gr-qc/0311082.

[40] H. Casini, M. Huerta, "A finite entanglement entropy and the c-theorem", Phys. Lett. B **600**, 142 (2004), arXiv:hep-th/0405111v3.

[41] T. Shimomura, T. Okamura, T. Mishima, H. Ishihara, "Does the generalized second law hold in the form of time derivative expression?", Phys.Rev. D **62** 044036 (2000), arXiv:gr-qc/9902029v1.

[42] N. Pippenger, "The inequalities of quantum field theory", IEEE Trans. on Info. Theory, **49**, 773 (2003).

[43] A.N. Bernal, M. Sánchez, "Globally hyperbolic spacetimes can be defined as 'causal' instead of 'strongly causal'", Class. Quant. Grav. **24**, 745 (2007), arXiv:gr-qc/0611138v1.

[44] A. Strominger, D. Thompson, "A Quantum Bousso Bound", Phys. Rev. D **70**, 044007 (2004), arXiv:hep-th/0303067.

[45] R.P. Geroch, "Topology in General Relativity", J. Math. Phys. **8**, 782 (1967).

[46] N. Sakai, K. Nakao, H. Ishihara, M. Kobayashi, "Is it possible to create a universe out of a monopole in a laboratory?", Phys. Rev. D **74**, 024026 (2006), arXiv:gr-qc/0602084v3.

[47] B. Freivogel, V.E. Hubeny, A. Maloney, R.C. Myers, M. Rangamani, S. Shenker, "Inflation in AdS/CFT", JHEP 0603**,** 007 (2006), arXiv:hep-th/0510046v4.

[48] S. Coleman, F. De Luccia, "Gravitational effects on and of vacuum decay", Phys. Rev. D **21,** 3305 (1980); W. Fischler, D. Morgan, J. Polchinski, "Quantum nucleation of false-vacuum bubbles" Phys. Rev. D **41,** 2638 (1990); E. Farhi, A.H. Guth, J. Guven, "Is it possible to create a universe in the laboratory by quantum tunneling?", Nucl. Phys. B **339,** 417 (1990); E.I. Guendelman, J. Portnoy, "The Universe out of an Elementary Particle?", Class. Quant. Grav. **16,** 3315 (1999), arXiv:gr-qc/9901066v1.

[49] T. Banks, "Heretics of the False Vacuum: Gravitational Effects On and Of Vacuum Decay 2", arXiv:hep-th/0211160v2; A. Aguirre, M.C. Johnson, "Dynamics and instability of false vacuum bubbles", Phys. Rev. D **72,** 103525 (2005), arXiv:gr-qc/0508093v2.

[50] G. Galloway, K. Schleich, D. Witt, E. Woolgar, "The AdS/CFT Correspondence Conjecture and Topological Censorship", Phys. Lett. B **505**, 255 (2001), arXiv:hep-th/9912119v2.

[51] M. Alcubierre, "The warp drive: hyper-fast travel within general relativity", Class. Quant. Grav. **11,** L73 (1994), arXiv:gr-qc/0009013; Allen E. Everett, "Warp drive and causality", Phys. Rev. D **53,** 7365 (1996); M.J. Pfenning, L.H. Ford, "The unphysical nature of 'Warp Drive' ", Class. Quant. Grav. **14,** 1743, (1997), arXiv:gr-qc/9702026; W.A. Hiscock, "Quantum effects in the Alcubierre warp drive spacetime", Class. Quant. Grav. **14,** L183 (1997), arXiv:gr-qc/9707024; C. Van Den Broeck, "A 'warp drive' with more reasonable total energy requirements", Class. Quant. Grav. **16,** 3973 (1999), arXiv:gr-qc/9905084; F.S.N. Lobo, M. Visser, "Fundamental limitations on 'warp drive' spacetimes", Class. Quant. Grav. **21,** 5871 (2004), arXiv:gr-qc/0406083; F.S.N. Lobo, "Exotic solutions in General Relativity: Traversable wormholes and 'warp drive' space-times", *Classical and Quantum Gravity Research*, ed. M.N. Christiansen, T.K. Rasmussen, Nova Sci. Pub. 2008, arXiv:0710.4474; S. Finazzi, S. Liberati, C. Barceló, "Semiclassical instability of dynamical warp drives", Phys. Rev. D **79,** 124017 (2009), arXiv:0904.0141.

[52] R.D. Sorkin, E. Woolgar, "A Causal Order for Spacetimes with C^0 Lorentzian Metrics: Proof of Compactness of the Space of Causal Curves", Class. Quant. Grav. **13,** 1971 (1996), arXiv:gr-qc/9508018v3.

[53] E. Witten, "Instability of the Kaluza-Klein vacuum", Nucl. Phys. B **195,** 481.

[54] D.N. Page, S. Surya, E. Woolgar, "Positive Mass from Holographic Causality", Phys. Rev.Lett. **89,** 121301 (2002), arXiv:hep-th/0204198.

[55] R.D. Sorkin, D. Sudarsky, "Large fluctuations in the horizon area and what they can tell us about entropy and quantum gravity", Class. Quant. Grav. **16,** 3835 (1999), arXiv:gr-qc/9902051.

[56] A. Vilenkin, Phys. Rev. D **27,** 2848 (1983); A. Linde, "Eternally existing self-reproducing chaotic inflanationary universe", Phys Lett. B **175,** 395 (1986); A. Linde, D. Linde, A. Mezhlumian, "From the Big Bang Theory to the Theory of a Stationary Universe" Phys. Rev. D **49,** 1783 (1984), gr-qc/9306035.

[57] M. Niedermaier, M. Reuter, "The Asymptotic Safety Scenario in Quantum Gravity", Liv. Rev. Relativity, **9,** 5 (2006), http://www.livingreviews.org/lrr-2006-5.

[58] P. Wallden, "Causal Sets: Quantum gravity from a fundamentally discrete space-time", J. Phys. Conf. Ser. **222**, 012053 (2010), arXiv:1001.4041v1.

[59] H. Sahlmann, "Loop quantum gravity – a short review", arXiv:1001.4188v1.

[60] F. Hacquebord, "Symmetries and interactions in matrix string theory", arXiv:hep-th/9909227v1.

[61] S.L. Dubovsky, S.M. Sibiryakov, "Spontaneous breaking of Lorentz invariance, black holes and perpetuum mobile of the 2nd kind", Phys. Lett. B **638,** 509 (2006), arXiv:hep-th/0603158; C. Eling, B.Z. Foster, T. Jacobson, and A.C. Wall, "Lorentz violation and perpetual motion", Phys. Rev. D **75,** 101502(R) (2007), arXiv:hep-th/0702124; T. Jacobson, A.C. Wall, "Black Hole Thermodynamics and Lorentz Symmetry", Found. Phys. **40,** 1076 (2010), arXiv:0804.2720; D. Blas, S. Sibiryakov, "Horava gravity vs. thermodynamics: the black hole case", arXiv:1202.4497; P. Berglund, J. Bhattacharyya, D. Mattingly, "Mechanics of universal horizons", arXiv:1202.4497. See also the explicit use of Lorentz invariance in Ref. [29].

[62] S. Coleman, "Black holes as red herrings: Topological fluctuations and the loss of quantum coherence", Nucl. Phys. B **307,** 867 (1988).

Conclusion

Therefore, the Universe Has a Cause

9

Libertarian Agency and the Craig/ Grünbaum Debate about Theistic Explanation of the Initial Singularity

J. P. Moreland

Though some would demur, it is widely recognized today that the spatio-temporal physical universe had a beginning. There are at least two classical Big Bang models of the initial singularity of time t=0. The first model depicts a time interval which is closed at t=0. On this model, t=0 is a singular, temporally first event of physical space-time. The second model features a time interval which is finite, open in the past, and excludes t=0 as a point of space-time. Rather, t=0 is construed as a boundary of space-time. In both models, there is no instant of time prior to the initial singularity. Now, a number of theists have seized on the recognition that the universe had a beginning to revive the ancient *kalām* cosmological argument for God's existence. The defender of the *kalām* cosmological argument presents three exhaustive dilemmas and tries to establish the relevant horn of each: (1) Either the universe had a beginning or it did not. (2) If the universe had a beginning, then it was either caused or uncaused. (3) If the beginning of the universe was caused, the cause was either personal or impersonal.

Needless to say, the *kalām* argument has been hotly contested. Let us grant that the past is temporally finite and that the universe had a beginning. In recent years, William Lane Craig and Adolf Grünbaum have engaged in a stimulating and heated debate about dilemma (3).[1] To understand this exchange, let us begin by asking this question: If the origin of the universe is not to be taken as a brute fact, what sort of explanation should we give for it? For at least two reasons, Craig has argued that it is obvious that an event-causal explanation will not suffice.[2] First, whether t=0 is the first event in time or the boundary of time and therefore not an event, it is the case that (i) if causal priority entails temporal priority, then there can be no event-cause for t=0, and (ii) if causal priority does not entail temporal priority (that is, a cause can be simultaneous with its effect), then the event cause for t=0 would itself be the (explanatorily) first event, rather than the initial singularity, and an event-causal explanation would in turn be needed for it, ad infinitum.[3] Second, Craig has claimed that the cause of the universe's beginning must be uncaused, eternal, and changeless (*sans* the creation). Moreover, the cause must be personal "[f]or the only way in which

a temporal effect could originate from an eternal, changeless cause would seem to be if the cause is a personal agent who eternally chooses to create an effect in time. A changeless, mechanically operating cause would produce either am immemorial effect or none at all; but an agent endowed with [libertarian] free will can have an eternal determination to operate causally at a (first) moment of time and thereby to produce a temporally first effect.[4]

On Craig's view, then, not only is an event-causal explanation inadequate for the beginning of the universe, but a personal explanation is adequate. Grünbaum has inveighed against a personal explanation of the beginning of the universe and raised a number of objections against Craig that seek to show the unintelligibility or inadequacy of a theistic personal explanation for the initial singularity. The purpose of this article is to show that once the conceptual resources of libertarian agency and agent causation are clarified, these resources are sufficient to rebut the objections raised by Grünbaum against Craig and to tip the scales of their limited debate in favor of Craig. I will not seek to justify the *kalām* cosmological argument in its entirety, nor will I argue that libertarian agency and agent causation are, in fact, the correct model of human action, though I believe that to be the case. I will begin by discussion event and agent causation and the forms of explanation appropriate to each. This will be followed by an application of agent causation and personal explanation to the task of rebutting the five key arguments proffered by Grünbaum against Craig.

Event Causation, Agent Causation and Explanation: Event causation is a model of efficient causality widely employed in science. Suppose a brick breaks a glass. In general, event causation can be defined in this way: an event of kind K (the moving of the brick) in circumstances of kind C (the glass being in a solid and not liquid state) occurring to an entity of kind E (the glass object itself) causes an event of kind Q (the breaking of the glass) to occur. Here, all causes and effects are events that constitute causal chains construed either deterministically (causal conditions are sufficient for an effect to obtain) or probabilistically (causal conditions are sufficient to fix the chances for an effect to obtain). Associated with event causation is a covering law model of explanation according to which some event (the explanandum) is explained by giving a correct deductive or inductive argument for that event. Such an argument contains two features in its explanans: a (universal or statistical) law of nature, and the initial causal conditions.

Some philosophers—compatibilists, for example—describe human actions in terms of event causality and employ a covering law model to explain such actions. Advocates of libertarian freedom demur and they have developed different versions of an alternative model of human action. A widely accepted understanding of libertarian agency identifies the core component of intentional action as intentional endeavoring, that is, exercising a power in trying to bring about some effect for a reason.[5] As we will see later, some libertarians offer slightly modified versions of intentional action. But for now, we may incorporate this characterization of intentional action in the following depiction of libertarian agency: Person P exercises libertarian agency, and freely and intentionally brings about some event e just in case 1) P is a substance that has the active power to bring about e; 2) P exerted his power as a first mover (an "originator")

to bring about e; 3) P had the categorical ability to refrain from exerting his power to bring about e; 4) P acted for the sake of a reason which serves as the final cause or teleological goal for which P acted. Taken alone, 1–3 state necessary and sufficient conditions for a pure voluntary act. Propositions 1–4 state necessary and sufficient conditions for an intentional act.

By "substance" I mean a member of a natural kind, an essentially characterized particular that sustains absolute sameness through (accidental) change and that possesses a primitive unity of parts, properties, and capacities/powers at a time. "Active power" is an epistemically primitive notion which has a sense that is ultimately understood ostensively in acts of first person introspective awareness of one's own initiation of change.[6] A characteristic mark of active power is the ability to initiate motion, to bring something about. Active power is a dual ability. So understood, it is impossible for an exercise of active power to be causally necessitated by prior events. A "first mover" is a substance which has active power. The notion of "categorical ability" in 3) has two important aspects to it. First, it expresses the type of ability possessed by a first mover that can exercise active power and, as such, it contrasts with the conditional ability employed by compatibilists. Second, categorical ability is a dual ability; if one has the ability to exert his power to do (or will to do) A, then one also has the ability to refrain from exerting his power to do (or to will to do) A. Finally, 4) expresses a view of reasons as irreducible, teleological goals for the sake of which a person acts. In general, we may characterize this by saying that person S Φ' d (for example, went to the kitchen) in order to Ψ (for example, get coffee or satisfy S's desire for coffee). This characterization of action, according to 4), cannot be reduced to a causal theory of action that utilizes belief/desire event causation.

Three things should be mentioned about this definition of libertarian agency. First, there are two basic schools of thought regarding the issue of whether or not agents cause their actions. Advocates of the first school hold to agent causation and, thus, believe that the first mover in 2) causes his actions. For example, Roderick Chisholm claims that an adequate theory of agency should include the notion that "there is a certain event or state of affairs that an agent caused or brought about ('Jones killed his uncle' implies that the death of Jones' uncle was caused or brought about by Jones . . .").[7] Along similar lines, Marilyn Adams states that "[w]e adults with impaired freedom are responsible for our choices . . . in the sense that we are the **agent causes** of them."[8] Advocates of the second school accept a non-causal view of agency in which the actions of unmoved movers are uncaused events done for reasons as final causes. Such reasons provide explanations for free acts but they do not do so by citing the reason itself or the agent as an efficient cause of the act. Stewart C. Goetz is a major advocate of the non-causal theory of libertarian agency: "On the non-causal view of agency I am developing, causation plays no **positive** role in explaining the reality . . . of freedom. . . . Actions with respect to which an agent is free and responsible are uncaused events."[9] Either way, libertarians agree that an unmoved mover is an agent that can act without sufficient causal conditions necessitating that the agent act—the agent is the absolute source of his own actions.

Second, libertarian agency theorists are divided about the role or reasons in an overall theory of agency. Non-causal theories of agency are clear in seeing reasons as final causes—teleological goals for the sake of which someone acts. Advocates of agent causation either accept this view of reasons or else they hold reasons to be necessary (efficient) causal conditions that, together with the agent's own active exercise of power (and, perhaps, other conditions), cause the action.[10] Therefore, some agent causationists would adjust 4) accordingly. Third, it is broadly logically impossible for a person to be caused to agent-cause something. Libertarian acts are spontaneous in the sense that there are no causal antecedents sufficient to determine that an agent act with libertarian freedom.

Advocates of libertarian agency employ a form of personal explanation that stands in contrast to a covering law model. To understand this form of explanation, we need to look first at a distinction that is part of action theory: the difference between a basic and non-basic action. To grasp the difference between a basic and non-basic action, note first, that often more than one thing is accomplished in a single exercise of agency. Some actions are done by doing others; for example, I perform the act of going to the store to get bread by getting into my car and by driving to the store. Basic actions are fundamental to the performance of all others but are not done by doing something else. In general, S's Φ-ing is basic if and only if there is no other non-equivalent action description 'S's Ψ-ing' such that it is true that S Φ-ed by Ψ-ing. My endeavoring to move my arm to get my keys is a basic action. A non-basic action contain basic actions are parts of and means to the ultimate intention for the sake of which the non-basic action was done. To fulfill a non-basic intention, I must form an action plan: a certain ordered set of basic actions that I take to be an effective means of accomplishing my non-basic intention. The action plan that constitutes going to the store to get bread includes the acts of getting my keys and walking to my car.[11]

In my view, an action is something contained wholly within the boundaries of the agent. Thus, strictly speaking, the results of an action are not proper parts of that action. A basic result of an action is an intended effect brought about immediately by the action. If I successfully endeavor to move my finger, the basic result is the moving of the finger. Non-basic results are more remote intended effects caused by basic results or chains of basic results plus more remote intended effects. The firing of the gun or the killing of Lincoln are respective illustrations of these types of non-basic results.

With this in mind, a personal explanation (divine or otherwise) of some basic result R brought about intentionally by person P, where this bringing about of R is a basic action A, will cite the intention I of P that R occur and the basic power B that P exercised to bring about R. P, I, and B provide a full explanation of R: agent P brought about R by exercising power B in order to realize intention I as an irreducibly teleological goal. To illustrate, suppose we are trying to explain why Wesson simply moved his finger (R). We could explain this by saying that Wesson (P) performed an act of endeavoring to move his finger (A) in that he exercised his ability to move (or will to move) his finger (B) intending to move the finger (I). If Wesson's moving his finger was an expression of an intent to move a finger to fire a gun to kill Smith, then we can explain the non-basic results (the firing of the gun and the killing of Smith) by saying that Wesson

(P) performed an act of killing Smith (I_3) by endeavoring to move his finger (A) intentionally (I_1) by exercising his power to do so (B), intending thereby to fire the gun (I_2) in order to kill Smith. An explanation of the results of a non-basic action (like going to the store to get bread) will include a description of the action plan.[12]

Personal Explanation, Divine Action, and Grünbaum's Critique: We are now in a position to discuss personal explanation and divine action.[13] Theists differ about the precise nature of the world's own causal activity, but I want to set these disputes aside and concentrate on explaining the difference between primary causal, miraculous acts by God in the natural world vs. the use of secondary causes by God.[14] In the ordinary course of natural events, weather patterns or chemical reactions, for example, God sustains natural entities in existence along with their own causal powers, but those entities have causal dispositions to bring about changes themselves if they are affected in certain ways. In ordinary causal chains, God does not move natural entities, he simply sustains them. In such cases, God is not directly responsible for the changes of states of affairs in things, and event causation, along with a covering law model of explanation, is appropriate, even if God employs natural causal sequences to secure a divine intent. However, when it comes to primary causal, direct miraculous acts by God in producing changes in the world (parting the Red Sea, creating the universe from nothing, for example), God exercises libertarian agency as a primary cause. Some have conceived of such activity as God either bringing something into existence *ex nihilo* or momentarily supplying natural entities with new causal powers different from their normal powers which, in turn, sets off a chain reaction of divergence in the universe. Irrespective of the precise nature of Divine primary causal activity, however, at least the basic results of such acts are caused by an exercise of Divine libertarian agency and are to be explained by personal theistic explanation.

From what I can tell, Adolf Grünbaum's rebuttal of Craig's employment of personal theistic explanation for the initial singularity boils down to five key arguments:

(1) The notion of a Divine agent cause of the initial singularity is incoherent because causation is essentially a temporal activity or relation in which the cause must be temporally prior to its effect. Now, if God were the agent cause of the initial singularity, then since there was no time prior to that singularity, this would mean that His creative causal act and the physical effect, that is, the initial singularity, would have to occur simultaneously. But the notion of simultaneous causation is unintelligible. Says Grünbaum, "I consider the notion of simultaneous causation, as applied to the purported **creation of time,** either unintelligible or, at best, incoherent."[15] Elsewhere, he adds, "Precisely the hypothesis that t = 0 simply had no temporal predecessor obviates the misguided quest for the elusive cause."[16]

(2) The notion of a Divine agent cause of the initial singularity is too inscrutable, mysterious and obscure to do any explanatory work. According to Grünbaum, any appeal to Divine agency as a creator, conserver, or intervener is an appeal to "the inherently irremediable dynamical inscrutability of divine causation . . ."[17]

He also claims that "so far as divine causation goes, we are being told . . . that an intrinsically elusive, mysterious agency X inscrutably produces the effect."[1818]

(3) Divine Personal explanation is inferior to scientific explanation because the former fails to provide what the latter proffers: a specification of intermediate causal processes, mediating causal links that connect causes (Divine or otherwise) with their effects. Says Grünbaum, ". . . the invocation of a divine creator to provide causal explanations in cosmology suffers from a fundamental defect vis-à-vis scientific explanation: As we know from two thousand years of theology, the hypothesis of divine creation does not even envision, let alone, specify, an appropriate intermediate causal process that would link the presence of the supposed divine (causal) agency to the effects which are attributed to it. . . . In physics, there is either an actual specification or at least a quest for the mediating causal dynamics linking presumed causes to their effect."[19]

(4) Grünbaum argues that it does not follow from the causal premise "whatever begins to exist has a cause" that the first cause is a conscious agent. So even if we grant that there is a first cause of the initial singularity, such an admission does not entail that the first cause is a personal God.[20]

(5) Only events can qualify as the momentary effects of other events or of the actions of agents and if the singularity t=0 is taken as a non-event, then the singularity cannot be taken as an effect of any cause: "Let me take for granted the altogether reasonable view that only events can qualify as the momentary effects of other events, or of the action of an agency. Since the Big Bang singularity is technically a non-event, and t = 0 is not a bona fide time of its occurrence, the singularity cannot be the effect of any cause in the case of either even-causation or agent-causation alike."[21]

Part of Craig's response to Grünbaum, especially to (1), is the specification of three cogent models of theological creationism of which two are relevant to my concerns. Model 1: The Creator may be conceived to be causally, but not temporally, prior to the origin of the universe, such that the act of causing the universe to begin to exist is simultaneous with its beginning to exist. The Creator, *sans* the universe, exists changelessly and (most likely) timelessly and at the singularity creates both the universe and time. Model 2: The Creator may be conceived to exist timelessly and to cause tenselessly the origin of the universe at the Big Bang singularity.

Further reflection on the nature of libertarian agency will, I hope, support the intelligibility of Craig's models and provide further refutations of Grünbaum's objections. For the purpose of brevity, let us assume for now without argument a volitional theory of action according to which normal actions, like raising one's arm, are to be parsed in the following way: the bodily movement is caused by a specific sort of event—a volition (endeavoring)—which, in turn, is brought about in some way or another by the agent himself.

Now there is a certain difficulty for the libertarian theory of agency if we grant that whatever has a beginning has a cause. The raising of one's arm is an event with a beginning and it is caused by another event—a volition. But the volition is an event

with a beginning and it has a cause as well, namely, the agent. Now, what does the agent do to cause his volition? If the agent does something, is what he does itself an event and if so, does it need a cause? The claim that the agent causes an infinite hierarchy of events in causing his volition is inadequate. What other solutions are available? There are three which have been most widely recognized.[22]

AC I: The agent does not do anything to cause his volition. The volition is a basic act produced directly by the agent without exercising any power to produce it. The agent is simply the first relatum that stands in a primitive causal relation to the second relatum, the volitional event. The following objection has been raised against AC I: If the volition occurs at a particular time t^1 and the cause is an enduring substance that existed for some time prior to t^1, then why did the volition occur when it did? One reason this problem arises is that in cases of libertarian agency, no set of conditions within an agent is sufficient to produce a volition. There may be necessary conditions (motives, beliefs, and desires, for example), but these may exist in an agent over a protracted time period with no volition brought about. If the agent doesn't do something to cause the volition, why does it happen at t^1? So far as I can see, short of abandoning AC I, the best solution to the problem is to work with the second relatum. An agent does not just cause a volition simpliciter, say a volition-to-ϕ. The agent causes a volition-to-ϕ-at-t_1 (or now).

AC II: The agent does do something to cause his volition, namely, he exercises a power. According to this view, the causal relation between an agent and his volition is not primitive; it is grounded in an exercise of power. In AC II, we should revise the causal principle and recognize that an exercise of power is not an event in the sense relevant to this revised principle. The causal principle should read "every substance that begins to exist or every change that a substance undergoes has a cause." Now an exercise of power is simply the exertion of a self-moving power or principle of self-determination that is not itself a change undergone by the agent. In libertarian acts, agents are unmoved or first movers. They do not first undergo a change (an exercise of power) before they can cause a change (a volition). Rather, agents, *qua* substances, directly cause their volitions by virtue of possessing and exercising their power to do so. Since an exercise of power is not a change undergone by an agent (nor a coming-to-be of a substance), it is not an event with a beginning in the sense relevant to the causal principle, even though there was a time before and after which the agent caused his volitions. Besides coming into existence, only changes (internal or relational) need a cause.

AC III: The correct causal principle is "every event that can broadly logically have a cause does have a cause." As we saw earlier, it is broadly logically impossible for someone to be caused to agent-cause something else, a volition, for example. So if we grant that an exercise of power is an event (that is, a change within the agent), when we recognize that such an exercise just is the event of an agent directly agent-causing his volition (the exercise of power isn't an event caused by the agent which, in turn, event causes the volition), it becomes clear that it does not have an efficient cause because it cannot (though it may have a reason which serves as the final cause of the exercise of power).

Let us apply these insights to Grünbaum's criticisms of divine agent causality. It turns out that his objections actually take exception to agent causation and not divine action *per se*. Regarding objection (1), the notion of divine agent causality is not incoherent at all. There is a well-developed literature about agent causality, as well as its application to divine action, and most philosophers who reject agent causation do not do so on the grounds that it is incoherent. Moreover, Craig's first response to the charge of incoherence (God's act of causing the origin of the universe was simultaneous with its beginning) is quite plausible. God's volition (AC I-III) or his exercise of power and his volition (AC III) could be taken as simultaneous or coincident with the basic result of God's volition—the creation of the initial singularity. AC I-III render intelligible the notion that a timeless God *sans* creation could spontaneously bring about the initial singularity. On AC I, God timelessly causes the volition-to-create-the-universe-at-t=0 and this volition is simultaneous or coincident with the initial singularity. AC II depicts God's exercise of power as something brought about by God but which is not itself a change within God. So it becomes intelligible to suppose that God could changelessly exist and exercise His power to bring about a volition to cause the-creation-of-the-universe-at-t=0. Again, the volition would be simultaneous or coincident with the initial singularity. If God could do this changelessly, I see no reason to deny that he could also do it timelessly because temporality is relevant to efficient causality largely because it is sometimes mistakenly thought that an exercise of efficient causality involves change, and change is temporal. On AC III, God's exercise of power would just be the uncaused event of directly agent-causing his volition to bring about the initial singularity. Such an exercise would not, indeed could not be preceded by a prior event that brought it about. In this sense, such an exercise is done spontaneously by an agent.

Craig's second model (God exists timelessly and tenselessly causes the initial singularity) is plausible as well (setting aside problems of reconciling this model with an A series view of time) if we set aside the volitional theory of action: an agent brings about an effect without doing so by way of an internal, volitional change within the agent. On AC I, God could timelessly cause from eternity the-beginning-of-the-universe-at-t=0. On AC II, an exercise of power would bring about the initial singularity but the exercise of power would not itself be an event. AC III would not be applicable to Craig's second model because it depicts an exercise of power as an event within the agent.

Objection (2) is wide of the mark as well. For one thing, the Divine creation of the initial singularity is precisely analogous to human libertarian acts, for example, both involve first movers who initiate change. There is nothing particularly mysterious or inscrutable about the latter; so in the absence of some good reason to think that there is some specific problem with the initial Divine creation, the charge of inscrutability is question-begging. Moreover, we understand exercises of power primarily from introspective awareness of our own libertarian acts, and we use the concept of action so derived to offer third-person explanations of the behavior of other human persons. There is nothing obscure about such explanations for the effects produced by other finite persons, and I see no reason to think that this approach is illicit in the case of Divine initial creation. In fact, naturalists like John Searle, John Bishop, and Thomas

Nagel all admit that our basic concept of action itself is a libertarian one.[23] Searle goes so far as to say that our understanding of event causality is conceptually derived from our first-person experience of our own causation. There is a major tradition in philosophy that agent causation is clearer and more basic than event causation, and it may actually be that if any sort of causation is inscrutable, it is event causation. By claiming that God created the initial singularity, we mean that (1) there was no sufficient causal antecedent for the initial singularity (temporal or otherwise) apart from either (i) God directly causing the initial singularity without first doing something (AC I) or (ii) an exercise of Divine power to bring about the initial singularity (AC II or III); (2) in a basic libertarian act, God freely and spontaneously brought about a volition which caused t=0 as a basic result.

Objection (3) is also wide of the mark. A basic libertarian act that produces a basic result does not have intermediate causal links between the act and the result. The basic result is brought about directly. Moreover, the sort of personal explanation that goes along with basic libertarian actions and basic results does not employ intermediate causal links in its explanans. As we saw earlier, examples of this sort of explanation makes reference to the agent himself, the agent's intention, and the basic power exercised. Since theists depict the creation of the initial singularity as a basic Divine act in which a First Mover brought about an initiation of change, yielding the initial singularity as a basic result, there is no causal intermediary. So it can hardly count as a liability that a theistic personal explanation of the initial singularity fails to cite an intermediate causal process disallowed by the theistic model itself.

In philosophy of science, it is widely recognized that, given two competing paradigms, sometimes one rival will consider a phenomenon basic and not in need of a solution, empirical or otherwise. It may, therefore, disallow questions about how or why that phenomenon occurs and, thus, can hardly be faulted for not being fruitful in suggesting lines of empirical research for causal processes whose existence is not postulated by the theory. As Nicholas Rescher has pointed out,

> [o]ne way in which a body of knowledge S can deal with a question is, of course, by **answering** it. Yet another, important different, way in which S can deal with a question is by disallowing it. S **disallows** [Q] when there is some presupposition of Q that S does not countenance: given S, we are simply not in a position to raise Q.[24]

Not only is it not a vice that a theistic personal explanation fails to cite an intermediate causal process; it is also a virtue that the theistic explanation requires no such causal process. Whatever caused t=0 could not have done so by way of an intermediate temporal causal process of the type Grünbaum heralds as a virtue of scientific explanation since t=0 is either the first event or the initial boundary from which events ensue. Either way, there are no events prior to t=0 and, therefore, an event causal explanation that cites a causal process leading up to t=0 is out of the question. Thus, the theistic model is superior to Grünbaum's scientific model as an explanation of the initial singularity because the former eschews but the latter requires an intermediate causal linkage.

In my view, this third objection is nothing more than a question-begging claim that if agent causal explanations do not conform to those proffered by event causal explanations, then the former are inferior as general explanatory strategies. So understood, Grünbaum's third objection is not really directed primarily at theistic explanation. Instead, it is an attack on libertarian agency and the sort of explanation associated with it. It would seem, then, that this third objection is not very impressive. Some philosophers may think that compatibilist models of freedom are to be preferred to libertarian models, but it is simply question-begging to fault the latter on the grounds that they do not conform to modes of explanation that only become relevant if we have already eschewed libertarian action and personal explanation in favor of event causation and explanation.

Regarding (4), so far as we know, only conscious agents endeavor or will anything. So even if it is logically possible that the First Cause was not conscious, it is not epistemically possible in light of what we know about the types of agents who are capable of libertarian action. If the origin of the spatio-temporal universe is not to be taken as a brute fact, then as we have seen, Divine libertarian agency theory provides an explanation for the initial singularity and event causal explanations can provide no such explanation. Now, I know of no example of an endeavoring libertarian agent which was not conscious. Grünbaum must do more than just assert the bare logical possibility that the First Cause was not conscious. He must show that it is epistemically plausible to think that such an agent could exercise a power and will the initial singularity into being and yet not be conscious. And this Grünbaum has not done.

Further, John Bishop has argued that given the nature of contemporary scientific naturalism, "the idea of a responsible agent, with the 'originative' ability to initiate events in the natural world, does not sit easily with the idea of [an agent as] a natural organism."[25] Elsewhere, Bishop notes that ". . . the problem of natural agency is an ontological problem—a problem about whether the existence of actions can be admitted within a natural scientific perspective. . . . [A]gent causal-relations do not belong to the ontology of the natural perspective. Naturalism does not essentially employ the concept of a causal relation whose first member is in the category of person or agent (or even, for that matter, in the broader category of continuant or 'substance'). All natural causal relations have first members in the category of event or state of affairs."[26] If we grant that abstract objects do not have efficient causal powers, then Bishop's remarks, if correct, would seem to imply that the efficient cause of the initial singularity is not only conscious, but also supernatural.

Finally, in light of what we have seen in our analysis of libertarian agency and personal explanation, objection (5) is simply false or question-begging. Agents bring about their own exercises of power and, in virtue of our analysis of AC II, these may reasonably be taken as non-events. Given AC II, Grünbaum's claim that only events can qualify as the momentary effects of other events or actions of agents is not correct. Moreover, in non-basic actions, agents are first movers who generate a chain of events. If we take an exercise of power or a volition to be the first event in such a chain, the agent brought about this event. However, if we take a non-basic action to be the initiation of a temporal, causal sequence to follow, such that the exercise of power/

volition forms an initial boundary for that sequence, the agent still brought about the exercise of power/volition. Insofar as libertarian agents are first movers, when they perform non-basic actions, the causal sequence of events brought about by such agents have an initial boundary in a way exactly analogous to the origination of the spatio-temporal universe. In the case of normal libertarian non-basic actions, the agent brings about the initial temporal boundary of the sequence of events, as well as the rest of the ensuing chain. In principle, then, there is no difficulty with the idea of a Divine libertarian agent bringing about t=0, as well as the rest of the events that follow. So even if we grant that t=0 is a non-event, libertarian action theory has the resources to allow that it could be an effect of a free, Divine, creative act.

In sum, in their exchange about the adequacy of personal theistic explanation for the initial singularity, neither Craig nor Grünbaum clarifies the nature of libertarian agency and agent causation in order to shed light on their exchange. I have tried to advance this debate by doing just that. From what we have seen, irrespective of one's view about the adequacy of the *kalām* cosmological argument in general, Grünbaum's arguments against Craig's employment of personal theistic explanation fail to be persuasive.

Notes

1 William Lane Craig, "The Origin and Creation of the Universe: A Reply to Adolf Grünbaum," *British Journal for the Philosophy of Science* 43 (1992): 233–40; "Cosmology and Big Bang Cosmology," *Philosophia naturalis* 31 (1994): 217–24; "Prof. Grünbaum on Creation," *Erkenntnis* 40 (1994): 325–41; "A Response to Grünbaum on Creation and Big Bang Cosmology," *Philosophia naturalis* 31 (1994): 237–49; Adolf Grünbaum, "The Pseudo-Problem of Creation in Physical Cosmology," *Philosophy of Science* 56 (1989): 373–94; "Creation as a Pseudo-Explanation in Current Physical Cosmology," *Erkenntnis* 35 (1991): 233–54; "Some Comments on William Craig's 'Creation and Big Bang Cosmology'," *Philosophia naturalis* 31 (1994): 225–36.

2 I am assuming here that there is no metaphysical time prior to physical time.

3 If the initial singularity is taken to be an event, then the event cause would be simultaneous with the initial singularity in the sense of being at the same time as the singularity. If the initial singularity is taken as a boundary of time, then the cause, strictly speaking, would not be an event (in which case, event causality would not be applicable), but could still occur coincidentally in that both the cause and the effect occur at t=0. For more on this, see William Lane Craig, "Creation and Big Bang Cosmology," *Philosophia naturalis* 31 (1994): 222.

4 Craig, "Cosmology and Big Bang Cosmology," 219.

5 This is not quite correct because, sometimes, an agent performs an intentional action by simply allowing a certain sequence of events to take place. Here an agent gives a sort of passive permission and does nothing to stop a sequence of events that accomplish the agent's intent. The agent refrains from endeavoring.

6 Timothy O'Connor holds that the primitive, core element in our concept of causality is that of the "production" or "bringing about" of an effect. See Timothy O'Connor,

"Agent Causation," in *Agents, Causes, & Events,* ed. by Timothy O'Connor (New York: Oxford University Press, 1995), 175–78. For O'Connor, this core element is a genus of which agent and event causation are distinct species. My claim that active power is epistemically primitive is consistent with O'Connor's claim in this way. Through introspective awareness of my own free acts, I am aware of my own active power. By reflection, I may then form a concept of active power. By reflection on the concept of active power, I may then form a concept of causal production *per se* by removing from my concept of active power the component of being able to refrain.

7 Roderick M. Chisholm, "The Agent as Cause," in *Action Theory,* ed. by Myles Brand and Douglas Walton (Dordrecht, Holland: D. Reidel, 1976), 199.

8 Marilyn McCord Adams, "The Problem of Hell: A Problem of Evil for Christians," in *Reasoned Faith,* ed. Eleonore Stump (Ithaca: Cornell University Press, 1993), 313.

9 Stewart C. Goetz, "A Noncausal Theory of Agency," *Philosophy and Phenomenological Research* 49 (1988): 311; see also his "Libertarian Choice," *Faith and Philosophy* (forthcoming). Compare Alan Donagan, *Choice: The Essential Element in Human Action* (New York: Routledge & Kegan Paul, 1989); Richard Taylor, *Action and Purpose* (Englewood Cliffs, NJ: Prentice-Hall, 1966).

10 Cf. Randolph Clarke, "Toward a Credible Agent-Causal Account of Free Will," *Noûs* 27 (1993): 191–203.

11 There is some debate about whether each of these basic actions requires its own intending. Richard Swinburne argues that in performing actions which take a long time (writing a chapter), we do not exercise a separate volition for each intentional action (for example, willing to write the first sentence) that is part of the long-term act. Rather, we just intend to bring about the long-term effect by bringing about a generally conceived series of events and the body unconsciously selects a particular routine to accomplish that effect. See Richard Swinburne, *The Evolution of the Soul* (Oxford: Clarendon Press, 1986), 94–95. I leave the matter open except to note that to the degree that a non-basic action contains sub-acts of a discontinuous nature (picking up keys, getting into a car vs. a series of steps in taking an hour long walk), then it is more likely that sub-intentions are required to characterize adequately those sub-acts.

12 Thus, we see that there are at least three kinds of intentional actions: Basic actions with a basic intent (simply intentionally moving my finger), basic actions with non-basic intents (ultimate intents that have other intents as means, for example, intentionally squeezing my finger to fire a gun to kill Smith), and non-basic actions (those that contain sub-acts—sub endeavorings and intendings—as parts, for example, going to the store to buy bread).

13 For excellent treatments of divine action relevant to this article, see the chapters by Jonathan L. Kvanvig and Hugh J. McCann, Philip L. Quinn, Peter van Inwagen, and William P. Alston in *Divine & Human Action: Essays in the Metaphysics of Theism* (Ithaca, N.Y.: Cornell University Press, 1988).

14 The main views of the world's causal activity in relationship to God's are the full secondary causality view (God sustains the world in existence but in the normal course of things, the entities of the world exert their own causal powers and such exertions are sufficient to produce changes in the world), occasionalism (there are no autonomous, distinct causal powers possessed by created objects; God is the only true cause and no effect in nature is brought about by natural entities), and concurrentism (every event

cause has God collaborating with the natural causal entity, cooperating with its causal activity by ratifying that activity). My own view is the first position.

15 Grünbaum, "Creation as a Pseudo-Explanation in Current Physical Cosmology," 244. Cf. "Some Comments on William Craig's 'Creation and Big Bang Cosmology,'" 231–35.

16 Grünbaum, "Creation as a Pseudo-Explanation in Current Physical Cosmology," 239.

17 Ibid., 235.

18 Ibid., 235. On the same page, Grünbaum confesses "I, for one, draw a complete **explanatory** blank when I am told that God created photons."

19 Ibid., 234–35.

20 Grünbaum, "The Pseudo-Problem of Creation in Physical Cosmology." Compare William Lane Craig, "The Origin and Creation of the Universe: A Reply to Adolf Grünbaum," *British Journal for the Philosophy of Science* 43 (1992): 233–35.

21 Grünbaum, "Some Comments on William Craig's 'Creation and Big Bang Cosmology,'" 237.

22 William Rowe, *Thomas Reid on Freedom and Mortality* (Ithaca, N.Y.: Cornell University Press, 1991), 30–40, 145–61.

23 In Searle, *Minds, Brains and Science* (Cambridge, Mass.: Harvard University Press, 1984), 98; John Bishop, *Natural Agency* (Cambridge: Cambridge University Press, 1989), 58, 69, 72, 95–96, 103–104, 126–127, 140–141, 144; Thomas Nagel, *The View From Nowhere* (New York: Oxford University Press, 1986), 110–37.

24 Nicholas Rescher, *The Limits of Science* (Berkeley: University of California Press, 1984), 22.

25 *Natural Agency*, 1. Bishop's own solution eschews libertarian agency in favor of a version of the causal theory of action.

26 Ibid., 40. An interesting implication of Bishop's view is that naturalism cannot allow for there to be a first event in the absolute sense of not being preceded by other events because all events are caused by prior events or else they are simply uncaused. In the latter case, the coming to be of the event cannot be "natural" since it is just a brute fact. In the former case, this means that if the *kalām* cosmological argument is correct and there was a beginning to the universe, then the beginning itself was not a natural event nor was its cause if it had one. For more on this, see William Lane Craig and Quentin Smith, *Theism, Atheism, and Big Bang Cosmology* (Oxford: Clarendon, 1993).

10

Causation and the Logical Impossibility of a Divine Cause[1]

Quentin Smith

1 Introduction

Some interesting light is thrown on the nature of causation, the origin of the universe, and arguments for atheism if we address the question: Is it *logically* possible that the universe has an originating divine cause?

I think that virtually all contemporary theists, agnostics, and atheists believe this is logically possible. Indeed, the main philosophical tradition from Plato to the present has assumed that the sentence "God is the originating cause of the universe" does not express a *logical* contradiction, even though many philosophers have argued that this sentence either is synthetic and meaningless (e.g., the logical positivists), or states a synthetic and a priori falsehood (e.g., Kant and Moore), or states a synthetic and a posteriori falsehood (e.g., contemporary defenders of the probabilistic argument from evil).

I believe the prevalence of this assumption is due to the fact that philosophers have not undertaken the requisite sort of metaphysical investigation into the nature of causation. This investigation is the purpose of this paper; specifically, I shall argue that the thesis that the universe has an originating divine cause is logically inconsistent with all extant definitions of causality and with a logical requirement upon these and all possible valid definitions or theories of causality. I will conclude that the cosmological and teleological arguments for a cause of the universe may have some force but that these arguments, traditionally understood as arguments for the existence of God, are in fact arguments for the nonexistence of God.

2 Causal definitions and the notion of an originating divine cause

Something is a *continuing cause* of the universe if and only if it causes each state of the universe. Something is an *originating cause* of the universe if and only if it causes the earliest state of the universe. If time is continuous, "the earliest state" may refer to an

instantaneous state or (if the universe's history is half-open in the earlier direction) to a temporally extended state of some given length.

If big bang cosmology is true, the universe began to exist about fifteen billion years ago with the big bang. The big bang is the earliest state of the universe; "the big bang" may be taken to refer to a singularity that constitutes the first instantaneous state of the universe or (if one "cuts out" the singularity) to an explosion that constitutes the first half-open state of some brief length, e.g., the Planck length, 10^{-43} second. In my discussion, I shall treat the big bang as a logically possible example of an earliest state of the universe.

Considerations of agent causality are not germane to our discussion; our topic is the cause of the universe's beginning to exist, not the cause of God's *act of willing* that the universe begin to exist. We are not examining the relation between God (the agent) and his act of willing (the effect) but the relation between his act of willing (an event) and the beginning of the universe (another event). Thus, definitions of agent causality are irrelevant to our arguments; we are interested only in definitions of event causality, where the cause and the effect are both events.

Hume's definition of a cause

The most famous and influential definition of a cause is Hume's definition; indeed, most contemporary definitions include conditions that are similar in some respect to at least one of the three conditions included in Hume's definition:

> Contiguity in time and place is therefore a requisite circumstance to the operation of all causes. . . . Priority in time is . . . another requisite circumstance in every case. . . . [A] third circumstance [is] that of constant conjunction betwixt the cause and the effect. Every object like the cause produces always some object like the effect. Beyond these three circumstances of contiguity, priority, and constant conjunction I can discover nothing in this cause.[2]

Hume's definition includes three conditions for being a cause: temporal priority, spatiotemporal contiguity, and a nomological relation ("every object like the cause produces always some object like the effect").

(a) Temporal priority

If time began to exist with the universe, the "temporal priority" condition of Hume's definition implies that the universe cannot be caused to begin to exist, since there is no earlier time at which the cause could occur.

Even if there is time before the universe, the "temporal priority" condition rules out an originating divine cause if all divine acts are timeless.

However, the "temporal priority" condition only shows the universe cannot have an originating divine cause if time began to exist with the universe or if all divine acts are timeless. It is logically possible that time preceded the beginning of the universe,

even if there are no known laws of physics according to which the physical variable *t* can take values earlier than the time at which space and mass-energy began to exist. Further, it is logically possible that God exists in time and that a pre-universe time is occupied by God's mental life, which includes his volitions. Thus, it is logically possible for a divine volition to meet the "temporal priority" condition of Hume's definition. The intractable problems begin with the other two conditions.

(b) Spatiotemporal continguity

Hume's definition and many other definitions of causality require that the causal event is spatially in contact with, or is spatially near, the effect. God is said to be omnipresent, but this means that she is conscious of and stands in a volitional relation to each physical particular. It does not mean that divine volitions, which are nonphysical, touch or are in the spatial vicinity of the physical particulars that are the objects of these volitions.

God's act of willing that the big bang occur is not spatiotemporally contiguous with the big bang, since this act of willing does not have spatial coordinates. Two particulars *c* and *e* are spatiotemporally contiguous *only if* the spatial coordinates *x, y, z* that locate *c* on a manifold either are identical with the spatial coordinates *x′, y′, z′* of *e* or locate *c* in the neighborhood of *e*.

(c) Nomological relatedness

The third feature of Hume's definition, the nomological condition ("every object like the cause produces always some object like the effect"), is also common to many definitions of causality. Hume's definition belongs to the line of reductive definitions that define causes in terms of laws of nature and a set of noncausal relations (such as temporal priority and spatiotemporal contiguity) between two particulars *c* and *e*.[3] According to these definitions, *c* is a cause of *e* only if there is a law of nature *L* that enables a statement that *e* occurs to be deduced from the premises that *c* occurs and that the law *L* obtains. For example, Carl Hempel writes:

> [A] "cause" must be allowed to be a more or less complex set of circumstances or events, which might be described by a set of statements $C_1, C_2, \ldots, C_k$.
>
> Thus the causal explanation implicitly claims that there are general laws—let us say, $L_1, L_2, \ldots, L_k$—in virtue of which the occurrence of the causal antecedents mentioned in $C_1, C_2, \ldots, C_k$ is a sufficient condition for the occurrence of the explanadum event.[4]

A probabilistic law *L* may be permitted as well, in which case "to be deduced from" would be replaced by "to be inductively supported by."

However, the nomological condition for being a cause is logically inconsistent with a divine cause of the big bang, since God by definition is a supernatural being and his or her actions are not governed by laws of nature. Furthermore, the fact that God's willing is omnipotent makes "the big bang occurs" deducible from "God wills that the big bang occur" alone, without the need of any supplementary nomological premise,

which violates the condition that a nomological premise is a logically necessary condition for the derivation of the conclusion that the effect exists from premises one of which is that the causal event occurs.

At this point, we have already ruled out virtually every extant definition of causality, since most every definition includes either the spatiotemporal contiguity condition or the nomological condition. We are left with non-contiguity and singularist definitions of causality.

A noncontiguity definition does not mention spatiotemporal contiguity and does not require the cause to be both temporally and spatially contiguous with the effect; variants of noncontiguity definitions may allow for timeless divine acts and/or temporal divine acts that are not spatially near or in contact with the effect. A singularist definition allows an event to cause an effect in a single case, without the cause and effect needing to instantiate some law. However, the extant formulations that are singularist and/or noncontiguity definitions are few and far between and prove problematic for a defender of the logical possibility of an originating divine cause.

Ducasse's singularist definition of a cause

The most famous singularist definition of a cause is C. J. Ducasse's. Ducasse's conception

> defines the cause of a particular event in terms of but a single occurrence of it, and thus in no way involves the supposition that it, or one like it, ever has occurred before or ever will again. The supposition of recurrence is thus wholly irrelevant to the meaning of cause; that supposition is relevant only to the meaning of law.[5]

Since the nomological condition is explicitly rejected, it seems this definition applies to God's willing that the big bang occurs.

However, further inspection of Ducasse's definition shows that it does not apply, since his definition requires spatiotemporal contiguity. Ducasse claims that the cause *c* is a sufficient condition of the effect *e* and that *c* is sufficient for *e* if (*i*) *c* is a change that occurred during a time and throughout a space terminating at an instant *i* at a surface *s* of an object; (*ii*) the change *e* occurred during a time and through a space beginning at the instant *i* at the surface *s*; (*iii*) no change other than *c* occurred during the time and through the space of *c*; and (*iv*) no change other than *e* occurred during the time and through the space of *e*.[6] Thus, Ducasse's account meets the singularist criterion but not the noncontiguity criterion. (Although Ducasse calls his account a "definition" of a cause, it is only a partial definition, since he begins his definition with "if," not "if and only if.")

The transference definition of a cause

Another possible candidate for a singularist and noncontiguity definition is based on the transference definition of causation, offered by Hector-Neri Castañeda, Galen Strawson, David Fair, Jerrold Aronson, and others.[7] Castañeda states that "the heart of

production, or causation, seems, thus, to be transfer or transmission."[8] In the actual world, what is transferred is energy (according to Castañeda), but he uses the word "causity" as a generic term for whatever may be transferred. Can God's volition *transfer causity* to the big bang?

Castañeda's full theory implies a definition that includes the nomological condition: *c* is a cause of *e* if and only if (*i*) there is a transfer of causity from an object O_1 to an object O_2 in a circumstance *x,* with the event *c* being O_1's transmission of causity and the event *e* being O_2's acquisition of causity; (*ii*) every event of the same category as *c* that is in a circumstance of the same category as *x* is conjoined with an event of the same category as *e*.

Condition (*ii*) is intended as a nomological condition and thus rules out supernatural causes. But may we isolate (*i*), "the heart of causation," and successfully argue that a singularist, noncontiguity, and transference condition is satisfied by a divine volition? It appears not, since there is a problem with causity. The causity cannot be identical with energy (Castañeda's claim about the actual identity of causity), since there is no energy in God (God being nonphysical). Indeed, the causity cannot be anything physical, since God is nonphysical. Nor can the causity be anything nonphysical, since the big bang is wholly physical. Thus, there appears to be no viable candidate for the causity transferred.

Counterfactual definitions of causation

David Lewis's definition imports counterfactual conditions into the definition of causation and seems to lend itself to a noncontiguity and singularist conception. According to Lewis, *c* causes *e* if and only if (*i*) *c* and *e* are events and both occur and it is the case that either (*ii*) if *c* had not occurred, *e* would not have occurred, or (*iii*) there is a causal chain linking *c* and *e* and each link *d* in the chain is such that if *d* had not occurred, then *e* would not have occurred. Since there is no causal chain between a divine volition and the big bang, condition (*iii*) is inapplicable and we may concentrate on (*i*) and (*ii*).[9]

Are the divine volition and the big bang both events? According to Jaegwon Kim, an event is a substance exemplifying an *n*-adic property at a time.[10] Even if there is no pre-universe time, this need not rule out the applicability of Kim's definition to God's volition, since we may construe God's volition as simultaneous with the big bang. We may also follow Brian Leftow and allow that the logical position occupied by "at the time *t*" may be occupied by "at eternity" or "timelessly."[11] Alternately, we could follow Davidson and take an event as a particular that is not further definable, thus allowing that the divine volition is an event even if timeless.[12] Taking one of these routes, or, following Wolterstorff and others,[13] taking an event as something's exemplification of an *n*-adic property (without a time specification), would allow us to consider the divine volition as an event that is either timeless, simultaneous with, or earlier than the big bang. (In these various definitions, "event" and "state" may be taken to be synonyms.)

However, Lewis's counterfactual definition is not instantiated by a divine willing of the big bang. Let *c* be the divine willing of the big bang and let *e* be the big bang. If *e* had

not occurred, then *c* would not have occurred. But this implies the false proposition that *e* is the cause of *c,* since *c* is counterfactually dependent on *e.* In this case (to use Lewis's words about a problem he generally notes), "we have a spurious reverse causal dependence of c on e, contradicting our supposition that e did not cause c."[14]

Lewis solves this problem by denying the counterfactual "if *e* had not occurred, *c* would not have occurred." Lewis holds that it is instead true that "c would have occurred just as it did but would have failed to cause e."[15] But this entails that Lewis's definition cannot be instantiated by God's willing the big bang, since if *c* had occurred (if God had willed the big bang), then it would have necessarily caused *e* (the big bang); God is omnipotent and his willing is necessarily effective.

In summary, the above considerations suggest that there are no extant definitions of causality that are satisfied by God's willing the big bang to occur; I believe a survey of further extant definitions would show that most of them include at least one of the above-mentioned conditions (contiguity, a nomological condition, etc.), conditions that are violated by the divine volition. The definitions that do not include one of the above-mentioned conditions include some other condition that is violated by the divine volition; for example, John Mackie's definition of an INUS condition implies that a cause *c* is neither necessary nor sufficient for its effect *e* but is instead an insufficient and nonredundant part of an unnecessary but sufficient condition for *e.*[16] God's willing the big bang, however, is sufficient for the occurrence of the big bang and thus violates the condition "is neither necessary nor sufficient for its effect."

3 Causes and logically sufficient conditions

It may be responded at this juncture that the failure of God's creation of the big bang to satisfy any of the extant definitions of causality does not imply that God's volitional act is not a cause of the big bang. It may be that the correct definition of causality has not yet been discovered and that God's willing the big bang satisfies this correct, undiscovered definition. My argument that God cannot be a cause of the universe is at best a "weak inductive argument" based on the definitions that have been formulated up to the present point in time.

Further, the preceding considerations suggest a certain definition of causality that *is* satisfied by the originating divine volition, regardless of whether this definition has been defended by anybody. This definition reads: *c* is a cause of *e* if and only if *c* is a sufficient condition of *e* and *c* is earlier than *e.* This definition includes Hume's "temporal priority" condition but is both singularist and noncontiguous. (A definition is contiguous only if it includes both spatial and temporal contiguity.) This definition cannot be satisfied by an originating divine volition if all divine volitions are timeless or if there is no time before the beginning of the universe. But it is logically possible that there is time before the big bang and that a temporal deity performs a volition that both occurs before the big bang and is a sufficient condition of the big bang's occurrence.

It may also be said that we need not rely on the assumption that the divine volition must satisfy a *definition* of a cause in order to be a cause. It is arguable that causation is

a simple relation, a conceptual primitive, and thus that there is no definition that could capture its nature.[17]

These three responses to my discussion in part 2 are perhaps not unreasonable; indeed, at least the first response (that part 2 presents an "inductive argument" based only on extant definitions) contains some truth.

However, all three responses are unavailing in face of the following crucial fact: there is an entailment relation between *"c* is a cause of *e"* and *"c* is not a *logically* sufficient condition of *e."* It is the case that:

(1) For any two particular events or states *x* and *y,* if *x* is a logically sufficient condition of *y,* then *x* is not a cause of *y.*

For example, a body's being in motion is a logically sufficient condition of the body's occupying space, but the body's being in motion is not the cause of the body's occupation of space. However, God's willing that the big bang occurs is a logically sufficient condition of the big bang, for the propositions expressed by "God wills that the big bang occur" and "the big bang does not occur" are logically incompatible. The reason for this is that God is omnipotent and thus his willing is always successful (of logical necessity); if an omnipotent being *x* wills *e* and *e* does not occur, then *x* is not omnipotent—which is a contradiction. (God can do everything that is logically possible; God cannot create a stone that is too heavy for him to lift, but creating such a stone is not a logical possibility. God would never will something to occur if the occurrence of that something were logically impossible—God is omniscient and omnibenevolent and would not knowingly engage in any futile effort.)

The variables in proposition (1) range over particular events or states; they do not range over particular events taken together with laws of nature or universal generalizations under which the particulars are subsumed. As we have seen, the nomological definitions of deterministic causation imply that a particular event *c,* in conjunction with a law of nature, logically necessitates the event *e* that is the effect. The sun's shining on a stone, in conjunction with the law that whatever is shined upon is warmed, logically necessitates that the stone is warmed. Proposition (1), however, implies only that *the sun's shining on the stone* does not logically necessitate *the stone's being warmed.* The sun's shining on the stone is a nonlogically sufficient condition of the stone's being warm (it is *nomologically sufficient,* in that it is logically sufficient for the stone's being warm only if it is conjoined with some law of nature).

Two objections may be made to my argument that divine volitions are logically sufficient conditions and therefore are not causes.

Objection 1. It may be objected that every cause can be described in a way that logically implies the occurrence of its effect and, therefore, that divine volitions are *not* dissimilar to causes. For example, the cause *the explosion that burned down the house* logically necessitates its effect *the burning down of the house,* since it is a logical contradiction to assert that "there is an explosion that burned down the house and yet there is no event of the house burning down."

But this objection is fallacious since "the explosion that burned down the house" does not refer merely to the cause but also refers to the effect. A definite description that refers merely to the causal event can be satisfied consistently with the non-occurrence of the effect; for example, the definite description "the explosion that occurred in the house" can be satisfied consistently with the nonsatisfaction of "the burning down of the house."

The fallaciousness of this objection can be explained more precisely in terms of referentially transparent and referentially opaque contexts. The definite description "the explosion that resulted in the burning down of the house" is a referentially transparent context; this implies that "the burning down of the house" occupies a position in "the explosion that resulted in the burning down of the house" that is open to substitution and quantification. A description of the form "the explosion that resulted in the *F*" permits coreferring expressions to be substituted for "the *F*," and if a description of this form is satisfied, it follows that there is an *F*. Since the description of the effect, "the *F*," occurs in a referentially transparent context, "the explosion that resulted in the *F*" refers to both the cause and the effect.

By contrast, the definite description "the divine willing that the big bang occurs" is a referentially opaque context and refers merely to the divine volition. This description is referentially opaque since it is a propositional attitude construction, and positions within attitude constructions are not open to substitution and quantification.[18] Specifically, a definite description of the form "the willing by *x* that the *F* occurs" does not permit substitutions of coreferring expressions for "the *F*," and "the *F*" is not open to quantification. This implies that if a description of the form "the willing by *x* that the *F* occurs" is satisfied, it does not follow that there is an *F*. Since the description of the effect, "the *F*," occurs in an opaque context, "the willing by *x* that the *F* occurs" refers only to the cause.

Given this distinction, we may say that a definite description *D* of a cause also refers to the effect if and only if *D* includes a term for the effect that is open to substitution and quantification. A definite description *D′* of a cause does not refer to the effect if and only if *D′* either does not contain a term for the effect or contains a term for the effect in an opaque context.

This enables us to state our principle (1) about causes and logically sufficient conditions in semantic terms: The satisfaction of a definite description *D* of a cause logically implies the existence of the effect if and only if *D* includes a term for the effect in a referentially transparent context. Since the satisfaction of the definite description "the divine willing that the big bang occurs" logically implies that the big bang occurs, despite the fact that "the big bang" does not occur in a referentially transparent context, it follows that this description does not refer to a *cause*.

The reason that the satisfaction of the description "the divine willing that the big bang occurs" implies that there is a big bang is not due to the logical form of the description (the form is opaque) but is due to the content of the description. This content is distinctive in that it makes the relevant conditionals about the divine volition and the big bang *logical truths*. The expression "an omnipotent being" means, in part, a being whose acts of will necessarily actualize what is willed. Thus, the sentence "if an

omnipotent being wills that the big bang is actualized, then the big bang is actualized" expresses the same proposition as the sentence "if an omnipotent being, whose acts of will necessarily actualize what is willed, wills that the big bang is actualized, then the big bang is actualized," which is a truth of logic.

Principle (1) about causes and logically sufficient conditions implies that *no causal conditional is a logical theorem,* where a causal conditional has the form "if *c* occurs, then *e* occurs" and substitutions for *"c"* are expressions that refer to the cause and do not include a term for the effect in a referentially transparent context.

These restatements of principle (1) in semantic and logical terms suffice to refute the first objection to my argument that the divine volition is not a cause, the objection that "for each cause *c* and effect *e,* there is some description of *c* that logically implies the existence of *e*."

Objection 2. The first objection to my argument about causes and logically sufficient conditions was that divine volitions are not unique, since every cause can be described in a way that logically implies the effect. A second objection is that there is some description of God's willing the big bang to occur that does *not* logically imply that the big bang occurs, and therefore (for this different reason) divine volitions are not dissimilar to causes. The description "the willing that has for its aim the actualization of the big bang" can be used as a definite description of the relevant divine volition and yet "there occurs the willing that has for its aim the actualization of the big bang, but the big bang is not actualized" is not a logical contradiction. It follows (the objection goes) that God's willing need not be regarded as a logically sufficient condition of the big bang. The objector may argue that the existence of such descriptions implies that whether or not a divine volition logically necessitates the existence of its volitional object is not a fact about the divine volition itself but is relative to how the volition is described.

But this objection is invalid, since the existence of a description of the divine volition that does not logically imply that the big bang occurs is consistent with the divine volition's necessarily possessing the relational property of being conjoined with the occurrence of the big bang. This consistency is an instance of the more general principle that "something that necessarily possesses a certain property *F* can be described by a definite description *D* that does not include *F* among its descriptive conditions, and *D* will not imply that whatever satisfies *D* necessarily possesses *F*." For example, the number 9 necessarily possesses oddness and is described by "the number of planets in our solar system"; but since "the number of planets in our solar system" does not include oddness among its descriptive conditions, it does not imply that whatever satisfies this description necessarily possesses oddness.

These responses to the two objections help to justify my claim that the proposition

(1) For any two particular events or states *x* and *y,* if *x* is a logically sufficient condition of *y,* then *x* is not a cause of *y*

both is true and precludes divine volitions from being causes.

Sosa's theory of causation

Does every philosopher accept that a particular event *c* that causes a particular event *e* cannot logically necessitate *e*? Ernest Sosa has suggested a theory of causality that might appear to be inconsistent with this thesis. Sosa distinguishes several types of causation: nomological causation, material causation, consequentialist causation, and inclusive causation. Of interest to us is Sosa's definition of consequentialist causation, since this definition is instantiated by God's willing the big bang to occur. In cases of consequentialist causation, "the cause does entail the result or consequence."[19]

Sosa lists several examples of consequentialist causation: (*i*) an apple's being red causes the apple to be colored; (*ii*) Tom's being in the room causes the general fact that there is someone in the room; (*iii*) Peter, Paul, and Mary are tall and the only people in the room, and this causes the general fact that everyone in the room is tall; (*iv*) an apple's being sweet, juicy, etc., causes the apple to have the value of goodness.

Sosa acknowledges that he has no analysis or definition of consequentialist causation, but he says it involves a consequence deriving necessarily from a cause "that is somehow more basic."[20]

The immediate rejoinder to Sosa's theory is that his cases of consequentialist causation are not cases of causation but cases of logical derivation or, more exactly, cases where the instantiation of one property *F* logically necessitates the instantiation of a second property *G,* or cases where the obtaining of one fact *p* logically necessitates the obtaining of a second fact *q.* When Sosa says that consequentialist causation does "seem to be a genuine form of causation,"[21] he seems to be mistaken. Indeed, the man or woman on the street, contemporary philosophers and scientists would all emphatically and correctly assert that these are not genuine cases of causation. But in fairness to Sosa, he acknowledges this very point and makes some plausible observations:

> It might be objected that much of the foregoing is a mere terminological maneuver, that it simply takes what philosophers have long called causation, relabels it 'nomological causation', and goes on to classify it with certain wholly other relations that philosophers have not heretofore called causal relations. And it might perhaps be that the word 'cause' and its cognates have been so closely and so persistently associated with nomological causation by philosophers that they must be surrendered. But even then the basic point would remain, for nomological causation is a relation between a source and a consequence or result, and so is material causation (e.g. generation), so is consequentialist causation (e.g. the apple is chromatically colored as a result of being red) and so is inclusive causation. . . . These are all source-consequence relations or result-yielding relations.[22]

Thus, we can agree with Sosa that causation can be classified with other result-yielding relations—such as the logical necessitation of a property *F* by another property *G*—as one type of result-yielding relation, but at the same time we can distinguish causation from these other noncausal result-yielding relations.

4 Analogical and literal descriptions

I suggest that the foregoing considerations give us good reason to believe that there is no actual or possible correct theory or definition of causality that is instantiated by God's willing the big bang.

How might the defender of divine causality answer these arguments? One answer might be to grant that God's willing is not a "cause" of the universe's beginning but instead is the "creator" or "producer" of the universe's beginning. But this change in terminology does not solve the problem: *"c* creates *e"* and *"c* produces *e"* each imply *"c* causes *e,"* so the problem is not avoided. If we wish to stipulate that *"c* creates *e"* does not imply *"c* causes *e,"* then we deprive the word "creates" of any apparent intelligibility. If "creates" no longer means what it normally means, then we are hard put to say what it means.

A similar problem affects an alternative solution, namely, that we say that God "wills" the universe to begin to exist but does not "cause" it to begin to exist. I provisionally used the terminology of "God's willing" and "divine volition" in the preceding discussion, but this usage calls for reevaluation. A statement of the form *"x* wills *e* and *e* occurs because of *x*'s willing" logically implies *"x*'s willing causes *e."* If God's act of willing is *not* an act of causation, it is difficult to say what the word "willing" means when applied to God. It does not mean what it means in such sentences as "John moved his broken limb by a sheer act of will."

Perhaps we can say that the words "willing" and "cause" are used in an analogical or metaphorical sense when applied to God. This means that God has some features that are analogous to the features we normally ascribe by "willing" and "cause" and also that God has some features that are different. The analogy for "willing" would be this: If a human wills something, this willing is a mental event that has for its aim bringing another event into existence. Likewise, we may say of God that he or she experiences a mental event and that this mental event has for its aim bringing another event into existence. This is the analogy. There is also a difference, in that God's willing is a logically sufficient condition for the existence of the event that is willed, whereas a human's willing is not logically sufficient for the event that is willed.

However, this resort to the "analogical" use of "willing" and "cause" threatens to break down the intelligibility of our talk about God's willing. The explanation of the analogical meaning of these words is in terms of other words that also have an analogical meaning. We said that God's willing is a mental event that "has for its aim bringing another event into existence." However, the literal meaning of the phrase about aiming for a goal implies that "it is logically possible that this goal is not achieved." When we say that Alice has the aim of writing a book, we mean, in part, that it is logically possible that she not succeed in achieving her aim. Given the literal meaning of "aims," a statement of the form *"x* aims to realize *F* and *F* is realized" is neither a logical nor an analytic truth. Consequently, the explanation of the analogical meaning of "divine willing" in terms of "aiming to do something" cannot involve a literal use of "aiming to do something." But if "aiming" is used analogically, then our problem of explaining what we mean by our words reappears. This problem does not appear to have a solution; we are embarked on a regress of explaining analogically used

words in terms of other analogically used words, with no way to end this regress by an explanation that involves words in their normal and literal use. This regress is vicious; in order to understand phrase one, we need to understand phrase two, but in order to understand phrase two, we need to understand phrase three, and so on. This suggests that we cannot attach any definite meaning to the assertion that God causes, wills, or aims to bring the universe into existence.

A literal formulation of the divine relation to the big bang

But this is not to say that we cannot intelligibly talk about God and her relation to the big bang. It appears that we can at least say that there is some *n*-adic property *F* exemplified by God such that by virtue of exemplifying this property, God stands in relation to the big bang of being a logically sufficient condition of the big bang. Perhaps we can even be more precise and say that *F* is some mental property, where "mental" is understood in terms of intentionality (in the tradition of Brentano, Husserl, Chisholm, and Searle). Further, we can say this intentional act that is experienced by God has a certain property as its intentional object, the property *being the big bang*. The property *being the big bang* will thereby have a second-order property, viz., *being the intentional object of the divine intentional act A, such that being an intentional object of A is a logically sufficient condition of being exemplified*. Talk of "intentional act" may be literal here, since these are technical terms in the philosophical literature and "act" here has a different meaning than "act" in "Jane acted quickly to remedy the situation" or "the last act of the play was a disappointment."

If it is objected that "intentional act" does not have a univocal meaning between "humans perform (embodied, non-omniscient, and non-omnipotent) intentional acts" and "God performs (disembodied, omniscient, and omnipotent) intentional acts," then we can resort to a more general level of talk. We can say that there is a certain relation *R* in which God stands to the property *being the big bang* such that by virtue of God's standing in *R* to *being the big bang,* it is logically necessary that *being the big bang* is exemplified.

In summary, we are safe in saying that God does not cause the big bang but *R*s the big bang, where "God *R*s the big bang" means that God stands in a certain relation *R* to *being the big bang* such that by virtue of standing in this relation to this property, it is logically necessary that this property is exemplified. (For ease of expression, I will sometimes talk loosely in the following discussion of God's standing in *R* to the big bang, but such talk should be strictly analyzed in the way I analyzed "God *R*s the big bang.")

5 Objections to the arguments that God cannot be a cause

First objection

It may be objected that the divine relation *R* cannot merely be that of being a logically sufficient condition of the big bang. God's standing in this logical relation to the big

bang is not similar to the *sun's being orange* standing in relation to the *sun's being colored* as a logically sufficient condition. The sun's exemplification of being orange does not in any sense bring about or produce the sun's exemplification of being colored. But God's exemplification of *R* does bring about the big bang.

But this objection is overtly question-begging. I have already argued that God's standing in relation to the big bang does not satisfy any extant definition of causation (part 2) and that it does not satisfy a logically necessary condition of being a cause (part 3). Thus, to introduce synonyms of "causes," such as "brings about" or "produces," etc., is simply to beg the question at issue.

It may be countered by the objector that there is an important disanalogy between the case of the relevant divine event and the case of other logically sufficient conditions, viz., that God's standing in relation to the big bang is an event, *a concrete particular,* and that the big bang is another concrete particular, whereas the other logical relations are between *abstract objects.*

This countering argument is inaccurate. According to one conception of events or states, an event or state is the exemplification of a property by something. God's exemplification of the polyadic property *R* is a state, and so is Jane's exemplification of running and her exemplification of being alive. The concrete state of Jane's exemplification of running is a logically sufficient condition of the concrete state of Jane's exemplification of being alive. Thus, there are two concrete states standing in the relation of one's being the logically sufficient condition of the other. This situation is similar in this respect to God's standing in the relevant relation to the big bang.

Nonetheless, the intuition may persist that there is an important ingredient in God's relation to the big bang of logically necessitating it that is not present in the sun's orangeness logically necessitating the sun's being colored, or in Jane's running necessitating her being alive, an ingredient that is metaphorically captured by causal language (e.g., "produces," "brings about," etc.). The objector may simply state that it is intuitively obvious that there is this difference between the two cases, even if this difference cannot be adequately expressed in words.

But this amounts to retreating to an ineffability theory. We now have the theory: "God does not literally cause the big bang but in some metaphorical sense causes the big bang, even though it is impossible to specify literally the analogy between causation and God's relation to the big bang that justifies the metaphor." The ineffability theory is that God's *R*-ing the big bang is a relation with two properties: one property of God's *R*-ing the big bang is that the *R*-ing is a logically sufficient condition of the big bang, and the second property is an indescribable property, which we may call an *X*-property, such that the *X*-property is a property of God's *R*-ing that makes the *R*-ing analogous in a relevant respect to a causal relation.

However, the ineffability theory fails for three reasons.

First, if the *X*-property makes the *R*-ing analogous to a causal relation, then the *X*-property is some property shared in common by the causal relation and the *R* relation. Since the *X*-property belongs to the causal relation, and we can literally describe the causal relation, we should be able to literally specify the causal relation's *X*-property and say that it is this property that the *R* relation has in common with the causal relation. But the ineffability theory fails to do this.

Second, the ineffability theory has no justification for asserting that there is this *X*-property. The ineffability theory mentions no datum that the postulation of the *X*-property is used to explain, and it introduces no premises from which the presence of the *X*-property is deduced. The only apparent justification might be that one has had a mystical experience and has directly "beheld" God's *R*-ing the big bang and "beheld" the *X*-property of this *R*-ing, but that in reporting this intuition, one realized that there are no adequate and literally used words that could describe this *X*-property. However, if the theory that God metaphorically causes the universe to exist amounts to nothing more than dark sayings about what is "beheld" in an ineffable mystical experience, then this is not a theory based on natural reason but is a flight into mysticism and the deliverances of "supernatural reason." It would hold no interest for a philosopher intent on constructing a worldview based on natural reason.

Third, the best explanation of the origins of the "intuition" that God metaphorically causes the big bang, and is not merely a logically sufficient condition of the big bang, does not imply that this intuition is true. The origin of this "intuition" is the long and pervasive tradition (in philosophy, religion, and "ordinary language") of using causal words—"causes," "creates," "wills," etc.—to describe God's relation to the beginning of the universe. The psychological associations produced by the adoption of this linguistic tradition give rise to the "intuition" that there must be an *X*-property of God's relation to the big bang that grounds the metaphorical usage of "causes."

There *are* differences between (for example) the orange-color relation and the *R* relation of God to the big bang, but none are causal-like. Orange is a kind of color, but God's relation to the big bang is not a kind of big bang. Further, the orangeness is a monadic property of the same thing of which being colored is a property, but the *R* property is polyadic and interconnects different objects. Thirdly, being orange and being colored are both physical properties, whereas the divine relation is a mental property and *being the big bang* is a physical property.

We can also specify formal features of the *R* relation: it is asymmetric, transitive, and irreflexive, but many noncausal relations also possess these formal features.

Second objection

The theist, agnostic, or atheist who believes it is logically intelligible to say that God is an originating cause of the universe may take the bull by the horns and arrogantly assert that God's being a logically sufficient condition of the big bang is a *counterexample* to the extant definitions of causation discussed in part 2, showing that these definitions are wrong, and that it is also a valid *counterexample* to my principle (1) which states that causes are not logically sufficient conditions. The objector proclaims: "All actual and possible contiguity or nomological definitions of causation are false. The correct definition is a noncontiguity and singularist definition that allows that some causal relations are logical relations."

The problem with this "arrogant objection" is that there is no apparent justification *apart from God's alleged acts of causation* for the belief that there is a correct definition of causation that is noncontiguous, singularist, and permits logical relations. But these

acts of God are precisely the events whose causal nature is in dispute. To assume, in face of the arguments I have given, that these acts are causal relations is a question-begging response. In order to *demonstrate* that the relevant divine relation is a causal relation, we must have a logically independent reason to believe there is some correct definition of causation that the divine relation *R* satisfies. But there is no such reason. Consider the argument:

(2) There is a sufficient reason *J* to believe that there is a correct definition of causation that is singularist, noncontiguous, and permits logical relations.

Therefore,

(3) the divine relation *R* is a causal relation.

If the offered reason *J* is (3), then the argument that the divine relation *R* is a causal relation is question-begging.

It may be objected that the defender of the "there cannot be a divine cause" thesis is in a similar question-begging situation and thus that there is a stand off. It may be said that the defender begs the question by assuming that (3) is false or cannot play the role of reason *J*.

This objection fails since the defender of the "there cannot be a divine cause" thesis has a non-question-begging argument for the falsity of (3). The argument is that all cases of causation not in dispute are inconsistent with the hypothesis that there is a correct definition of the sort mentioned in (2). Both parties to the dispute agree that physical events cause other physical events and that the mental events of intelligent organisms cause other events (assuming an appropriate philosophy of mind), and this agreement is the common ground between the opponent and defender of the "there cannot be a divine cause" thesis. But this common ground is inconsistent with the positive thesis—viz., that there can be a divine cause—if only for the reason that it is a logically necessary property of the agreed-upon cases of causation that the causal event *is not* a logically sufficient condition of the effect. Since these causal events are necessarily *not* logically sufficient conditions, a definition of *a cause* that encompassed both these causal events and God's relation to the big bang would include the contradiction "is not a logically sufficient condition and is a logically sufficient condition." The agreed-upon cases may also include nomological and contiguity conditions, and consequently there may be further contradictions—e.g., "instantiates some law of nature and does not instantiate any law of nature" and "is spatially contiguous with the effect and is not spatially contiguous with the effect."

Third objection

It may be argued that a disjunctive definition can solve the problem. Suppose we have this disjunctive definition of causation: *c* is a cause of *e* if and only if either *c* is a logically sufficient condition of *e* or *c* is not a logically sufficient condition of *e* and instead satisfies (say) the Humean conditions.

One problem with this disjunctive definition is that it classifies the sun's being orange as a cause of the sun's being colored. So it does not work for this reason, as well as for the other reasons mentioned in my discussion of Sosa's account of causation.

Even if we add a temporal priority condition, this disjunctive definition will not work. We may say: *c* is a cause of *e* if and only if *either c* is both a logically sufficient condition of *e* and temporally prior to *e or c* is not a logically sufficient condition of *e* and satisfies (say) the Humean conditions. However, the first disjunct is satisfied by many items that are not causes. John's being a living organism (or John's *being embodied in a mortal body at time t*) both is temporally prior to and is a logically sufficient condition of John's being dead, but John's being a living organism (or John's *being embodied in a mortal body at time t*) is not the cause of his death. His death is caused, say, by a car hitting him as he crosses the street. The concept expressed by "is a living organism" analytically includes the concept expressed by "is mortal," and the relevant logical truths (e.g., "if *x* is an organism that dies, then *x* dies") can be obtained by substituting synonyms for synonyms.

Suppose we become even more specific and say instead: *c* is a cause of *e* if and only if *either c* is God's standing in the *R* relation to *e or c* is not a logically sufficient condition of *e* and satisfies (say) the Humean conditions. But this attempt to produce a satisfactory definition fails for two interrelated reasons.

First, a logically necessary condition of a correct definition of a purely qualitative *universal,* be it a monadic property or a relation (such as causation or intentionality), is that it not include a disjunct mentioning one particular case that does not meet the general conditions described in the other disjunct. A purely qualitative universal does not include any particulars as constituents. An example of an impurely qualitative universal is *being taller than Mount Everest.* Definitions of purely qualitative universals mention general conditions and do not include any mention of a particular case, such as the particular case of God's standing in an *R* relation to something.

Second, if this logical condition of correct definitions of purely qualitative universals (viz., the condition of not mentioning a particular case in a disjunct) were allowed to be violated, then the procedure of testing definitions by the counterexampling method (the standard method of testing the correctness of definitions) would no longer be usable. Any counterexample to a definition could be made consistent with the definition by adding to it a disjunct that mentions the counterexample. To save the definition *"x* is a planet if and only if *x* is a large body that orbits a star and contains no life," we can expand it to *"x* is a planet if and only if *either x* is a large body that orbits a star and contains no life *or x* is the Earth." The distinction between correct definitions and ad hoc definitions would collapse.

Fourth objection

A final argument is that philosophers from Plato to Plantinga have described God's relevant mental state as a cause of the universe and, therefore, this is an acceptable notion. There is both an established philosophical usage for calling God's relation to

the universe a "causal relation" and a long and venerable tradition that held it to be coherent to describe a divine mental state as a cause.

This argument, which is in effect an appeal to authority, is unsuccessful, since if this argument were admissible, it could be used to reject any new theory that is inconsistent with traditionally held theories. This appeal to authority at best motivates us to examine seriously the notion that God's mental states are causes, in deference to the fact that virtually all philosophers and laypersons have accepted this notion as logically unproblematic.

Perhaps to respond fully to this objection we also need an explanation of why this mistaken tradition has prevailed for so long and among so many philosophers. I think the main reason is that an investigation of the logical connection between what is expressed by "the universe's beginning to exist is the result of a divine act" and what is expressed by "the natural event *e* is the causal result of the natural event c" has not been systematically undertaken. (The main exception is the different but illuminating discussions of this connection in the recent writings of Adolf Grünbaum.)[23] Most philosophers have tacitly presupposed that the thesis that "divine causation is *logically* possible" is unproblematic; but once this thesis is examined, the presupposition is seen to be false.

6 Conclusion: Cosmological and teleological arguments for God's nonexistence

The argument of this paper might seem at first glance to tell us more about the nature of causation and the nature of God than about atheism versus theism. "A divine state cannot cause the universe to begin to exist" does not entail that God does not exist or that the big bang is not a logical result of a divine state. It merely entails that we cannot describe a divine state as the originating cause of the universe.

Nonetheless, there are important and perhaps decisive implications for the debate between theism and atheism, namely, that arguments from the necessary truth, a priori truth, or empirical truth of some causal principle cannot be a relevant premise from which to deduce or induce that the big bang is the logical consequence of God's standing in the relation *R* to the property *being the big bang*. Consider the following argument:

(4) Whatever begins to exist has a cause.

(5) The universe begins to exist.

Therefore,

(6) the universe has a cause.

This argument fails to support the thesis that God exists or that there is a divine cause of the universe. Indeed, this argument entails that the universe's existence is the result of

something other than a divine state, namely, a cause. Nor can any inductive argument that is based on the fact that every observed event has a cause be used to support the thesis that the big bang is the result of a divine state, since such an inductive argument instead would support the thesis that the big bang is the effect of some cause.

In fact, all the various cosmological and teleological arguments for the existence of God are really *arguments for God's nonexistence.* These arguments are arguments for the thesis that the universe has a cause and if the universe has a cause, God does not exist. This can be demonstrated as follows.

The traditional definition of God is: *x* is God if and only if *x* is omniscient, omnipotent, omnibenevolent, and is the cause of any universe that exists. We have seen that what is traditionally expressed by "God is the cause of the universe," if it is logically coherent, should be expressed instead by "God *R*s the universe." Thus the correct definition of God reads: *x* is God if and only if *x* is omniscient, omnipotent, omnibenevolent, and *Rs* any universe that exists. It follows from this definition that it is an essential property of God that he *R*s any universe that exists. Since this property is essential to God, there is no possible world in which it is true both that God exists and that there is a universe to which God does not have an *R* relation.

Our discussion of Sosa's theory of causation suggested that the causal relation and the divine *R* relation are two different types of *result-yielding* relations, to borrow Sosa's phrase. If the universe is the result of a *causal* result-yielding relation, it is not the result of an *R-type* result-yielding relation, and if the universe is the result of a divine act of *R*-ing, it is not the result of a cause. If there is a possible world in which some universe is the result of a cause, it follows that God does not exist in that possible world.

This shows how a cosmological argument for God's nonexistence may be explicitly constructed. The premises and inferences are set out in the following argument:

(4) Whatever begins to exist has a cause.

(5) The universe begins to exist.

Therefore,

(6) the universe has a cause.

(7) If the universe is the result of a cause, it is not the result of God's standing to the universe in an *R* relation.

(8) It is an essential property of God that he *R*s any universe that exists.

Therefore [from (7) and (8)],

(9) there is no possible world in which it is true both that God exists and that there is a universe which is the result of a cause.

Therefore [from (6) and (9)],

(10) God does not exist.

If big bang cosmology is true [and thus (5) is true], it seems that the premise with the weakest or lowest epistemic status is the first premise, "Whatever begins to exist has a cause." But William Lane Craig says about this premise: "[T]he first premiss is so intuitively obvious, especially when applied to the universe, that probably no one in his right mind *really* believes it to be false."[24] If Craig is right and my argument is sound, it follows that probably no one in his right mind who believes that the universe has a beginning *really* believes that God exists.

The same considerations apply to the teleological argument, one version of which reads:

(11) Artifacts are caused to exist by some intelligent being(s) with some purpose in mind.

(12) The universe resembles an artifact.

Therefore, it is probable that:

(13) The universe is caused to exist by some intelligent being(s) with some purpose in mind.

If this is an adequate argument from analogy, then it is probably true that the result-yielding relation that is involved in the explanation of why the universe exists is a causal relation in which some intelligent being(s) stand(s) to the universe. It follows [given propositions (7) and (9)] that God probably does not exist.

Since the cosmological and teleological arguments have standardly been thought to be the strongest arguments for God's existence, and since they support atheism rather than theism, it now seems that the case for theism is very weak indeed. It is hard to imagine how one could ever inductively or deductively establish, or find it self-evident, that the big bang is the logical consequence of something's standing in an *R* relation to *being the big bang.* Perhaps there are some fairly plausible arguments that the big bang has a cause, but there are no extant or plausible arguments that the big bang has a logically sufficient condition in an acausal mental state. This suggests that belief in the existence of God is considerably less reasonable than even the most cautious natural theologians have standardly supposed.

Notes

1 Earlier versions of this paper were read at West Virginia University (February 1995) and at Southern Methodist University (March 1996). The philosophers at both universities offered helpful comments on these earlier versions. Mark Aronszajn and William Lane Craig wrote critical responses to earlier versions, which proved useful in writing the present draft. I am also grateful to Christopher Hill for several suggestions that enabled me to improve an earlier draft.

Research for this paper was supported by an American Council of Learned Societies Fellowship for 1996 and by a National Endowment for the Humanities Summer Stipend for 1995.

2 David Hume, "An Abstract of a Treatise of Human Nature," in *An Enquiry Concerning Human Understanding* (New York: Bobbs-Merrill, 1955), 186–87.

3 It is worth noting that Michael Tooley's theory implies that a cause requires an underlying law of nature but that the cause is not specified solely by the law of nature and noncausal facts. Although Tooley's definition differs from the traditional *reductive* definitions, its inclusion of a nomological condition precludes it from being satisfied by a divine volition. See his *Causation: A Realist Approach* (Oxford: Clarendon Press, 1987).

4 Carl Hempel, *Aspects of Scientific Explanation* (New York: The Free Press, 1965), 348–49.

5 C. J. Ducasse, "On the Nature and the Observability of the Causal Relation," in *Causation,* ed. Ernest Sosa and Michael Tooley (Oxford: Oxford University Press, 1993), 129.

6 Ibid., 127.

7 Hector-Neri Castañeda, "Causes, Causity, and Energy," in *Midwest Studies in Philosophy IX,* ed. Peter A. French, Theodore E. Uehling, Jr., and Howard K. Wettstein (Minneapolis: University of Minnesota Press, 1984); Galen Strawson, "Realism and Causation," *Philosophical Quarterly* 37 (1987): 253–77; David Fair, "Causation and the Flow of Energy," *Erkenntnis* 14 (1979): 219–50; Jerrold Aronson, "The Legacy of Hume's Analysis of Causation," *Studies in the History and Philosophy of Science* 7 (1971): 135–36.

8 Castañeda, op. cit., 22.

9 See David Lewis, "Causation," *Journal of Philosophy* 70 (1973): 556–67.

10 Jaegwon Kim, "Events As Property Exemplifications," in *Action Theory,* ed. Myles Brand and Douglas Walton (Dordrecht: Reidel, 1976).

11 See Brian Leftow, *Time and Eternity* (Ithaca, N.Y.: Cornell University Press, 1993).

12 See Donald Davidson, "Causal Relations," in Sosa and Tooley, op. cit.

13 See Nicholas Wolterstorff, "God Everlasting," in *God and the Good,* ed. Clifton Orlebeke and Lewis Smedes (Grand Rapids, Mich.: Eerdmans, 1975); Quentin Smith, *Language and Time* (New York: Oxford University Press, 1993).

14 David Lewis, *Philosophical Papers,* vol. 2 (New York: Oxford University Press, 1983), 170.

15 Ibid.

16 See John Mackie, *The Cement of the Universe* (Oxford: Clarendon Press, 1974).

17 I argued for this approach in "The Concept of a Cause of the Universe," *Canadian Journal of Philosophy* 23 (1993): 1–24. In this earlier article, I claimed that cases of divine volitions are valid counterexamples to extant definitions of causality. However, I have since developed a counterargument to this claim (see part 5, the second objection) that has led me to abandon the thesis that divine volitions are causes.

18 More exactly, a *notational occurrence* of a term in a position within attitude constructions is not open to substitution and quantification; a *relational occurrence* of a term in this position is open to substitution and quantification. *"F"* occurs relationally in *"y* desires that there is an *F"* if this is read as "$(\exists x)$ *Fx. y* desires that: $(\exists x)$ *Fx*." By contrast, *"F"* occurs notationally if *"y* desires that there is an *F"* is read as *"y* desires that: $(\exists x)$ *Fx."* See David Kaplan, "Opacity," in *The Philosophy of W. V. Quine,* ed. Lewis Edwin Hahn and Paul Arthur Schilpp (La Salle, Ill.: Open Court, 1986). When I talk about "positions within attitude constructions" I have in mind only positions within attitude constructions in which terms occur notationally.

19 Ernest Sosa, "Varieties of Causation," in Sosa and Tooley, op. cit., 240.
20 Ibid.
21 Ibid.
22 Ibid., 242.
23 Adolf Grünbaum, "The Pseudo-Problem of Creation in Physical Cosmology," in *Philosophy and Physical Cosmology,* ed. John Leslie (New York: Macmillan, 1990), 92–112; "Creation As a Pseudo-Explanation in Current Physical Cosmology," *Erkenntnis* 35 (1991): 233–54.
24 William Lane Craig and Quentin Smith, *Theism, Atheism, and Big Bang Cosmology* (Oxford: Clarendon Press, 1993), 57.

11

Beyond the Big Bang

William Lane Craig

The Alternatives Before Us

The discovery that the universe is not eternal but had a beginning implies that the universe is not necessary in its existence and therefore has its ground in a transcendent being. The only way of avoiding this conclusion would be to deny Leibniz's conviction that whatever exists has a reason for its existence, either in the necessity of its own nature or else in an external ground, and to claim that the universe simply sprang into being uncaused out of nothing. Reflecting upon the current situation, P. C. W. Davies muses,

> What caused the big bang?' . . . One might consider some supernatural force, some agency beyond space and time as being responsible for the big bang, or one might prefer to regard the big bang as an event without a cause. It seems to me that we don't have too much choice. Either . . . something outside of the physical world . . . or . . . an event without a cause (Davies 1995: 8–9).

J. Richard Gott and Li-Xin Li seek to break this dilemma by defending the extraordinary hypothesis that *the universe created itself*. Observing that 'The question of first-cause has been troubling to philosophers and scientists alike for over two thousand years,' they note that modern scientists have, like Aristotle, found models of the universe attractive which involve the universe's sempiternal existence, since in this way 'one would not have to ask what caused it to come into being' (Gott and Li-Xin 1998: 023501–1). 'Now that it is apparent that our universe began in a big bang explosion', however, 'models with a finite beginning have taken precedence' and 'the question of what happened before the big bang arises' (Gott and Li-Xin 1998: 023501–1). They observe that inflation seemed to be 'a very promising answer, but as Borde and Vilenkin have shown, the inflationary state preceding the big bang could not have been infinite in duration – it must have had a beginning also. Where did it come from? Ultimately, the difficult question seems to be how to make something out of nothing' (Gott and Li-Xin 1998: 023501–1). Gott and Li-Xin, however, suggest instead that we should ask whether anything in the laws of physics would prevent the universe from creating itself.

Noting that General Relativity allows for the possibility of closed time-like curves, they hypothesize that as we trace the history of the universe back through an original inflationary state, we encounter a region of closed time-like curves prior to inflation. According to one possible scenario, a metastable vacuum inflates, producing an infinite number of (Big Bang type) bubble universes. In many of these a number of bubbles of metastable vacuum are created at late times by high energy events. These bubbles usually collapse and form black holes, but occasionally one will tunnel to create an expanding, metastable vacuum or baby universe. One of these expanding, metastable vacuum baby universes 'turns out to be the original inflating metastable vacuum we began with' (Fig. 11.1).

Gott and Li-Xin conclude that 'the laws of physics may allow the universe to be its own mother' (Gott and Li-Xin 1998: 023501–1).

Now we may leave it to the physicists to assess Gott and Li-Xin's claim that the laws of physics permit such a scenario, as well as the question of whether there are non-lawlike physical facts which contradict it. For the Gott-Li-Xin hypothesis raises fundamental metaphysical issues about the nature of time which, I think, render their hypothesis either metaphysically impossible or else superfluous.

Philosophers of time have distinguished two different views about the nature of time, which have been called the A- and the B-theories of time respectively.[1] According to the A-theory, temporal moments may be classed as past, present, and future, and only that moment which is present exists. Past moments and the things or events which occupy them have passed away and no longer exist; future moments, things, and events have not yet come to be and so do not yet exist. On the A-theory of time things

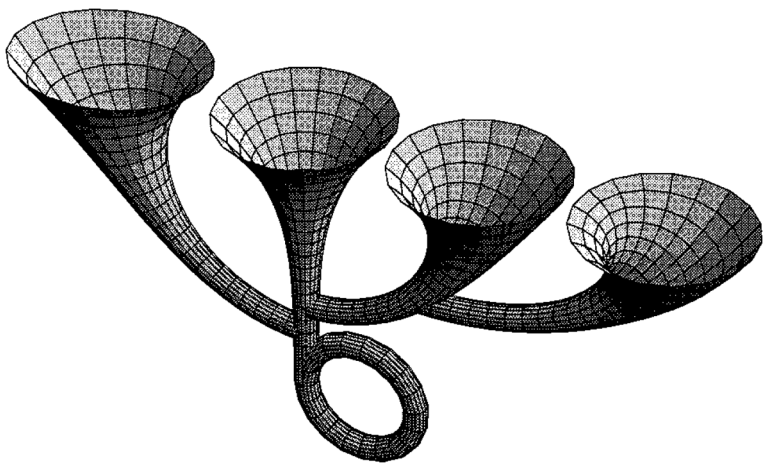

Fig. 11.1 A self-creating universe. Four inflating baby universes are shown. Universes A and D have not created any baby universes. Universe C has created universe D. Universe B has created three universes: A, C, and itself, B. The torus-shaped region at the bottom is a region of closed timelike curves. Such a universe neither arose from a singularity nor tunneled from noting, but it created itself.

come into and go out of being, and thus temporal becoming is a real and objective feature of reality.

By contrast, on the B-theory of time the distinction between past, present, and future is a subjective illusion of human consciousness. All things or events in time are equally real and existent, and moments, things, and events merely stand to one another in tenseless relations of *earlier than*, *simultaneous with*, or *later than*. Nothing ever comes into or goes out of being, and temporal becoming is an illusion.

Now all instances of causal influence over the past – whether we are talking about closed time-like curves, time travel, retro-causation, tachyonic anti-telephones, or whatever – presuppose the truth of the B-theory of time.[2] For clearly on the A-theory of time, at the time at which the effect is present, the cause is future and therefore literally non-existent. Thus, the effect just comes into being from nothing. Not only does this scenario seem absurd, but it also reduces to the first horn of Davies' dilemma with respect to the origin of the universe. The universe just came uncaused from nothing.

Thus the Gott-Li-Xin hypothesis presupposes the B-theory of time. But if one presupposes such a view of time, then Gott and Li-Xin's hypothesis becomes superfluous. For on a B-theory of time the universe never truly comes into being at all.[3] The whole four-dimensional spacetime manifold just exists tenselessly, and the universe has a beginning only in the sense that a meter-stick has a beginning prior to the first centimeter. Although the spacetime manifold is intrinsically temporal in that one of its four dimensions is time, nonetheless it is extrinsically timeless, in that it does not exist in an embedding hyper-time but exists tenselessly, neither coming into nor going out of being. The four-dimensional spacetime manifold is in this latter sense eternal. Thus, there is no need for the device of causal loops or closed time-like curves at the beginning to explain how it came into being.

Now space does not permit me to review the arguments for and against the A- and B-theories of time. I have explored this fascinating debate for the last dozen years and report my findings elsewhere.[4] Here I can only outline my reasons for affirming an A-theory of time as the most plausible view of the matter:

I. Arguments for the A-Theory
   A. Linguistic tense, which is ineliminable and irreducible, mirrors the tensed facts which are characteristic of reality.[5]
   B. The experience of temporal becoming, like our experience of the external world, should be regarded as veridical.[6]

II. Refutation of Arguments against the A-Theory
   A. McTaggart's Paradox is based upon the illicit assumption that there should exist a unique tenseless description of reality, as well as the illicit conflation of A-theoretic becoming with a B-theoretic ontology.[7]
   B. The passage of time is not a myth, but a metaphor for the objectivity of temporal becoming, a notion which can be consistently explicated on a presentist metaphysic.[8]

III. Refutation of Arguments for the B-Theory

A. Temporal becoming is wholly compatible with the mathematical core of Relativity Theory, even if its affirmation requires a different physical interpretation than the received view.[9]
B. Time, as it plays a role in physics, is an abstraction of a richer metaphysical reality, omitting indexical elements such as the 'here' and the 'now' in the interest of universalizing the formulations of natural laws.[10]

IV. Arguments against the B-Theory
A. In the absence of objective distinctions between past, present, and future, the relations ordering events on the B-theory are only gratuitously regarded as genuinely temporal relations of *earlier/later than*.[11]
B. The subjective illusion of temporal becoming involves itself an objective temporal becoming of contents of consciousness.[12]
C. The B-theory entails perdurantism, the view that objects have spatio-temporal parts, a doctrine which is metaphysically counter-intuitive, incompatible with moral accountability, and entails the bizarre counterpart theory of transworld identity.[13]

Given the truth of the A-theory of time, the idea that the universe is self-created, that is to say, brought itself into being via closed timelike curves, is either metaphysically impossible or else reduces to the notion that the universe sprang into existence uncaused out of nothing. Thus, I think that we are stuck with Davies' dilemma: the beginning of the universe is either an event without a cause or it is the result of a supernatural agency.

The supernaturalist alternative

Suppose we go the route of postulating some causal agency beyond space and time as being responsible for the origin of the universe. A conceptual analysis of what properties must be possessed by such an ultra-mundane cause enables us to recover a striking number of the traditional divine attributes. For as the cause of space and time, this entity must transcend space and time and therefore exist atemporally and non-spatially, at least *sans* the universe. This transcendent cause must therefore be changeless and immaterial, since timelessness entails changelessness, and changelessness implies immateriality. Such a cause must be beginningless and uncaused, at least in the sense of lacking any antecedent causal conditions. Ockham's Razor will shave away further causes, since we should not multiply causes beyond necessity. This entity must be unimaginably powerful, since it created the universe out of nothing.

Finally, and most strikingly, such a transcendent cause is plausibly to be regarded as personal. As Swinburne points out, there are two types of causal explanation: scientific explanations in terms of laws and initial conditions and personal explanations in terms of agents and their volitions (Swinburne 1991: 32–48). A first state of the universe cannot have a scientific explanation, since there is nothing before it, and therefore it can be accounted for only in terms of a personal explanation. Moreover, the personhood of the cause of the universe is implied by its timelessness and immateriality, since

the only entities we know of which can possess such properties are either minds or abstract objects, and abstract objects do not stand in causal relations. Therefore the transcendent cause of the origin of the universe must be of the order of mind. This same conclusion is also implied by the origin of a temporal effect from a timeless cause. For if the cause of the universe were an impersonal set of necessary and sufficient conditions, it could not exist without its effect. The only way for the cause to be timeless and changeless but for its effect to originate *de novo* a finite time ago is for the cause to be a personal agent who freely chooses to bring about an effect without antecedent determining conditions. Thus, we are brought, not merely to a transcendent cause of the universe, but to its personal creator.

Naturalistic objections

The naturalist, of course, will be reluctant to take on board such metaphysical baggage. But think of the alternative: that the universe came into being uncaused out of nothing. That seems metaphysically absurd. The naturalist philosopher of science Bernulf Kanitscheider remonstrates, 'If taken seriously, the initial singularity is in head-on collision with the most successful ontological commitment that was a guiding line of research since Epicurus and Lucretius,' namely, *out of nothing nothing comes,* which Kanitscheider calls 'a metaphysical hypothesis which has proved so fruitful in every corner of science that we are surely well-advised to try as hard as we can to eschew processes of absolute origin' (Kanitscheider 1990: 344). Mario Bunge thinks that an absolute origin of the universe 'would be unscientific, for science abides by the principles that nothing comes out of nothing or turns into nothingness . . . and that everything happens according to law rather than miracles' (Bunge 1985: 238–9). On the basis of the first principle Bunge, like Kanitscheider, rejects the view that the universe came into being uncaused out of nothing. On the basis of the second principle he thinks to reject theism. But while the principle that *out of nothing nothing comes* is a first principle of metaphysics as well as science, there is no incompatibility between being a theist metaphysically and a methodological naturalist scientifically; moreover, even methodological naturalism is far from unchallengeable.[14] It is difficult to see how any sensible person, particularly the naturalist, can think that the universe just sprang into existence uncaused out of nothing.

It has therefore been remarkable to observe in recent years the number of naturalists who, under the force of the evidence for an absolute beginning of the universe, have embraced the view that the universe is a surd contingent, something which popped into existence uncaused out of nothing. Quentin Smith declares, 'The fact of the matter is that the most reasonable belief is that we came from nothing, by nothing and for nothing'.[15] Rather than posit a cause of the origin of the universe, Smith advises, 'We should instead acknowledge our foundation in nothingness and feel awe at the marvelous fact that we have a chance to participate briefly in this incredible sunburst that interrupts without reason the reign of non-being' (Smith 1993a: 135).

Sometimes attempts are made to render this remarkable hypothesis more plausible, but these are usually not very impressive. Consider, for example, Peter Atkins's account of the origin of the universe:

> Now we go back in time beyond the moment of creation, to when there was no time, and to where there was no space . . . In the beginning there was nothing . . . . By chance there was a fluctuation, and a set of points, emerging from nothing, . . . defined a time . . . From absolute nothing, absolutely without intervention, there came into being rudimentary existence . . . Yet the line of time collapsed, and the incipient universe evaporated, for time alone is not rich enough for existence. Time and space emerged elsewhere, but they too crumbled back into their own dust, the coalescence of opposites, or simply nothing. Patterns emerged again, and again, and again. Each time the pattern formed a time, and through their patterning into time, the points induced their own existence . . . Sometimes chance patterned points into a space as well as a time . . . Then, by chance, there came about our fluctuation. Points came into existence by constituting time but, this time, in this pattern time was accompanied by three dimensions of space . . . with them comes stability, later elements, and still later elephants' (Atkins 1992: 129, 149–51).

This account is so obviously incoherent in postulating time before time and so confused in its reification of mathematical entities that we may rightly dismiss it as the pseudo-scientific drivel that it is.[16]

Or again, when John Gribbin asserts that the origin of the universe from nothing presents no problem, since the positive energy associated with mass is precisely offset by the negative energy associated with gravitation, so that in the case of the origin of the universe we got 'Not something for nothing, after all, but *nothing* for nothing' (Gribbin 1986: 374), he commits himself to the absurd position that nothing exists (not even he himself!). At the very best, the fact that the universe contains counter-balancing amounts of positive and negative energy could show that the universe need not have a material cause; but it does nothing to obviate the need for an efficient cause. As Isham puts it, there is still the 'need for ontic seeding' to produce the positive and negative energy, even if on balance it is naught (Isham 1994: 8). That is why the quantum vacuum was needed as a substratum in cosmogonic theories postulating such a process.

More often naturalistic thinkers have sought to commend their view either by attacking the causal principle *whatever begins to exist has a cause* or else by arguing for the implausibility or incoherence of the existence of a cause of the universe. Attacks on the causal principle are usually based on an appeal to quantum indeterminacy. For example, virtual particles are sometimes said to constitute a counter-example to the principle because they spring uncaused from the quantum mechanical vacuum. Wholly apart from the disputed question as to whether virtual particles really exist at all,[17] the central point to be understood here is that the quantum vacuum on which they depend for their existence is not nothing. It is for that reason that the statements frequently made with respect to Vacuum Fluctuation Models that 'the universe quantum tunneled

into being out of nothing,' or that 'nothingness is unstable' to fluctuations which grow into universes, or that 'the universe is a free lunch' because in this case 'we got something for nothing' cannot be taken seriously, for they treat nothing as though it were something, a sort of substance possessing properties and governed by the laws of quantum physics. In fact such statements turn out to be just rhetorical flourishes which no informed scientist takes literally. The quantum vacuum, which underlies all of space-time reality, is a fluctuating sea of energy. Because the vacuum is a physical entity existing in space and time, Vacuum Fluctuation Models did not envision a genuine origin of the universe out of nothing, as Kanitscheider emphasizes:

> The violent microstructure of the vacuum has been used in attempts to explain the origin of the universe as a long-lived vacuum fluctuation. But some authors have connected with this legitimate speculations [*sic*] far-reaching metaphysical claims, or at most they couched their mathematics in a highly misleading language, when they maintained 'the creation of the universe out of nothing' . . .
>
> From the philosophical point of view it is essential to note that the foregoing is far from being a spontaneous generation of everything from naught, but the origin of that embryonic bubble is really a causal process leading from a primordial substratum with a rich physical structure to a materialized substratum of the vacuum. Admittedly this process is not deterministic, it includes that weak kind of causal dependence peculiar to every quantum mechanical process (Kanitscheider 1990: 346–7).

Thus, quantum physics does not serve to rebut the principle that whatever begins to exist has a cause.

It is not surprising that naturalists should attack the notion of a cause of the universe, since they reject supra-natural realities independently of their motivation to justify an uncaused origin of the universe from nothing. Sometimes these critiques may be easily dismissed. For example, metaphysician John Post obviously begs the question when he claims that there cannot be a cause of the origin of the universe, since 'by definition the universe contains everything there is or ever was or will be' (Post 1991: 85). Again it is an obvious *non sequitur* when he infers that because 'the singularity cannot be caused by some earlier *natural* event or process,' therefore 'the universe has an uncaused beginning' and 'it seems contemporary physical cosmology cannot be cited in support of the idea of a *divine* cause or creator of the universe' (Post 1991: 87).

On the other hand, Smith realizes that the metaphysician must take seriously the 'more difficult question' of 'whether or not the singularity or the Big Bang probably is an effect of a supernatural cause, God' (Smith 1993a: 120). What problems, then, are there with a supernaturalist perspective? Adolf Grünbaum has argued vigorously against what he styles 'the New Creation Argument' for a supernatural cause of the origin of the universe.[18] His basic *Ansatz* is based on the assumption that causal priority implies temporal priority. Since there were no instants of time prior to the Big Bang, it follows that the Big Bang cannot have a cause.[19]

It seems to me that the supernaturalist has a number of options for dealing with this objection, one of which is to hold that the transcendent cause of the universe is causally, but not temporally, prior to the Big Bang event, such that His act of causing the universe to begin to exist is simultaneous, or co-incident, with its beginning to exist. Grünbaum provides no justification for his assumption that causal priority implies temporal priority. Discussions of causal directionality deal routinely with cases in which cause and effect are simultaneous. A supernaturalist could hold that the Creator *sans* the universe exists changelessly and, hence, timelessly and at the Big Bang singularity created the universe along with time and space. For the Creator *sans* the universe, there simply is no time because there are no events of any sort; time begins with the first event, at the moment of creation.

The time of the first event would be not only the first time at which the universe exists, but also, technically, the first time at which God exists, since *sans* the universe God exists timelessly.[20] The moment of creation is, as it were, the moment at which God enters time. His act of creation is thus simultaneous with the origination of the universe.

In response to this suggestion, Grünbaum has opposed the following argument:[21]

1. The proponent of simultaneous, asymmetric causation must furnish a generally accepted criterion for distinguishing one of two causally connected simultaneous events as the cause of the other, if simultaneous, asymmetric causation is possible.
2. There is no generally accepted account of causal directionality.
3. Therefore, there can be no simultaneous, asymmetric cause of the Big Bang.

The argument, if successful, would eliminate all purported instances of simultaneous, asymmetric causation, not just a cause of the Big Bang.

The argument, however, is, I think, unsound because (1) is so obviously false. For (i) Why must the proponent of simultaneous, asymmetric causation furnish a *generally accepted* criterion of causal directionality in order for such causation to be possible? Is this not an extravagant demand? Grünbaum fails to appreciate that there is no generally accepted account of causal directionality *überhaupt*, including accounts which appeal to temporal priority as a condition of causal priority. Indeed, I should dare to say that there is no generally accepted account of causation at all today. But should we therefore infer that causation is impossible or non-existent? Compare the situation in contemporary epistemology. There is today no generally accepted account of justification or rational warrant with respect to beliefs we hold to be true; but should we therefore infer that knowledge is impossible? Deconstructionists and other post-modernists may think so, but I doubt that Grünbaum would be ready to follow in their train. There is no reason to think that the possibility of simultaneous causation depends upon our being able to come up with an uncontroversial criterion of causal directionality. (ii) Indeed, what reason is there to think that the possibility of simultaneous, asymmetric causation depends upon my being able to come up with any kind of criterion of causal directionality at all? My enunciation of a criterion for distinguishing a cause from its effect is an epistemic affair; the existence of simultaneous

causation is a matter of ontology. A criterion helps us to *discern* simultaneous, asymmetric causes in the world; but to suggest that said criterion somehow *constitutes* such causal relations in reality is verificationism at its most implausible. Grünbaum has not suggested any incoherence or difficulty in simultaneous, asymmetric causation; if there are such causes in the world, they do not have to wait around for us to discover some criterion for distinguishing them. (iii) There is no reason to think that in order for specific cases of simultaneous, asymmetric causation to be possible or discernible, one must be able to furnish a general criterion broad enough to cover all such alleged cases. All one needs is a way of distinguishing cause from effect in the specific case. Now in the case of the hypothesis of theological creationism, we have a logically airtight means of distinguishing cause from effect, namely, it is *metaphysically impossible* for God to be caused by the world, since if God exists, His nature is such that He exists necessarily, whereas the world's existence is metaphysically contingent (as is evident from its beginning to exist). That entails that there is *no possible world* in which God is caused by the Big Bang. Hence, it is easy for the theist to explain in what sense God is causally prior to the universe or the Big Bang: God and the universe are causally related, and if the universe were not to exist, God would nevertheless exist, whereas there is no possible world in which the universe exists without God. Thus, it seems to me that Grünbaum's objection to a supernatural cause of the origin of the universe is unsuccessful.

The naturalist will perhaps raise a metaphysical objection to the scenario I have sketched of the Creator's status *sans* the universe. For it requires that we conceive of a timeless, personal agent, and some philosophers have argued that such a notion is self-contradictory.[22] For it is a necessary condition of personhood that an individual be capable of remembering, anticipating, reflecting, deliberating, deciding, and so forth. But these are inherently temporal activities. Therefore, there can be no atemporal persons.

The fallacy of this reasoning is that it conflates *common* properties of persons with *essential* properties of persons. The sorts of activities delineated above are certainly common properties of temporal persons. But that does not imply that such properties are essential to personhood. Arguably, what is necessary and sufficient for personhood is self-consciousness and free volition, and these are not inherently temporal. In his study of divine timelessness, John Yates writes,

> The classical theist may immediately grant that concepts such as reflection, memory, and anticipation could not apply to a timeless being (nor to any omniscient being), but this is not to admit that the key concepts of consciousness and knowledge are inapplicable to such a deity . . . there does not seem to be any essential temporal element in words like . . . 'understand,' to 'be aware,' to 'know,' and so on . . . an atemporal deity could possess maximal understanding, awareness, and knowledge in a single, all-embracing vision of himself and the sum of reality (Yates 1990: 173).

Similarly, God could possess a free, changeless intention of the will to create a universe with a temporal beginning. Thus, neither self-consciousness nor free volition entail

temporality. But since these are plausibly sufficient for personhood, there is no incoherence in the notion of a timeless, personal Creator of the universe.

More recently Smith has argued that 'the thesis that universe has an originating divine cause is logically inconsistent with all extant definitions of causality and with a logical requirement upon these and all possible valid definitions or theories of causality (Smith 1996: 169–70). Smith shows that the typical analyses of the causal relation in terms of temporal priority, spatio-temporal contiguity, and nomological relatedness are inapplicable to the event of God's willing that the Big Bang occur and the event of the occurrence of the Big Bang. Therefore, these two events cannot, on the customary analyses, be regarded as cause and effect. Counterfactual analyses of causation, such as David Lewis's, according to which *c* causes *e* iff (i) *c* and *e* are both events which occur and (ii) if *c* had not occurred, *e* would not have occurred, fare no better in Smith's view. For if *c* is God's willing and *e* is the Big Bang, it is true that if *e* had not occurred, then *c* would not have occurred. But this implies that the Big Bang is the cause of God's willing, which is false. Lewis avoids the problem of spurious reverse causal dependence by stipulating that if *e* had not occurred, then *c* would have occurred but failed to cause *e*. But since God is omnipotent and His willing necessarily effective, such a stipulation cannot be made in the present case. Thus, under no extant analysis of causality can God be said to cause the Big Bang.

Smith's argument may be formulated as follows:

4. If the claim that God caused the Big Bang cannot be analyzed in terms of extant definitions of causality, then God cannot have caused the Big Bang.
5. The claim that God caused the Big Bang cannot be analyzed in terms of extant definitions of causality.
6. Therefore, God cannot have caused the Big Bang.

Is this argument sound and persuasive? I think not.

Consider premise (4). I see no reason to think that this premise is true. In general, arguments to the effect that some intuitively intelligible notion cannot be analyzed in terms of current philosophical theories ought to make us suspect the adequacy of those theories rather than reject the commonsense notion. The idea that God caused the universe is intuitively intelligible. A cause is, loosely speaking, something which produces something else and in terms of which the thing that is produced can be explained. That notion certainly applies to God's causing the universe. Indeed, God's causing certain effects is analogous to our acting as agents to bring about certain effects. We certainly conceive of ourselves as causes, and, intuitively, God should count as a cause as well. But Smith's argument, if successful, could be generalized to prove that God is not a cause of anything whatsoever. If God's acting as a cause cannot be analyzed in terms of current philosophical definitions of causation, then so much the worse for those definitions! That only shows that the definitions need to be revised. Indeed, the standard procedure in terms of which proposed definitions of causality are assessed is to postulate counter-examples of intuitively obvious cases of causation and then show how the definition fails to accommodate these examples. In the same way, if God's being a cause cannot be accommodated by some philosophical definition

of causality, then that plausibly constitutes a counter-example to the definition which shows its inadequacy as a general metaphysical analysis of the causal relation, however adequate it might be for scientific purposes.[23]

Moreover, there is no reason to believe that we have arrived at the final and correct analysis of causation. In fact, there is good reason to believe the opposite. The definitions discussed by Smith are exclusively concerned with natural, even physical, causes. They were not intended to cover such recondite cases as divine causation of the origin of the universe. It is hardly surprising, therefore, that these analyses should fail to capture this notion. Smith simply ignores analyses of causation which are not currently fashionable but which were crafted in the context of a theistic metaphysic and are consonant with God's being the cause of the origin of the universe, for example, the account of efficient causation and creation given by Francisco Suarez in his monumental *Disputationes metaphysicae* (Suarez 2002). In his lengthy Introduction to his translation of questions 20–22 of Suarez's work, Freddoso argues that Suarez's account of causality not only enables one to construe God's creation of the universe as an instance of causation but also contrasts favorably with empiricist accounts of causality offered by contemporay philosophers such as Mackie, Lewis, van Fraassen, and Tooley.

Finally, Smith just assumes that an analysis of the causal relation can be given. But it could be held that such a relation is conceptually primitive, in which case we should not expect a successful reductive analysis to exist which will cover all cases. The plethora of competing extant analyses and the recognized deficiencies of all of them lend credibility to this viewpoint.

What about premise (5)? It seems to me that there are analyses of causation, however inadequate, which can accommodate God's causing the Big Bang. Consider Lewis's analysis of causation. According to Lewis, *c* causes *e* if and only if *c* and *e* are both events which occur and if *c* had not occurred *e* would not have occurred. Now God's willing the Big Bang clearly satisfies this definition: God's willing and the Big Bang are both events which occur, and if God's willing had not occurred, the Big Bang would not have occurred. But Smith rejoins, 'But if the Big Bang had not occurred, God's willing would not have occurred. So is the Big Bang the cause of God's willing?' Obviously not; but what this calls into question is the *adequacy* of Lewis's analysis, not whether divine causation satisfies it. Lewis remedies the problem by stipulating that if *e* had not occurred, *c* would still have occurred but failed to cause *e*, a remedy which will not work for divine causation. Actually Lewis's remedy will not work for many natural causes either, since in some cases the counterfactual, 'If *e* had not occurred, *c* would not have occurred' is true. So what Lewis's definition gives is not an analysis of '*c* causes *e*' but rather an analysis of '*c* and *e* are causally related,' and it fails to specify the *direction* of causation. But the theist faces no problem there: for, as we have said, it is metaphysically impossible for God's willing to have an external cause. There is no possible world in which the Big Bang causes God's volition. Therefore, given Lewis's analysis of '*c* and *e* are causally related' and the impossibility of the Big Bang's causing God's willing, it follows that God's willing causes the Big Bang. Thus, divine causation satisfies Lewis's definition of causality.

Again, there are analyses of agent causation which are even more relevant in the case of divine causation than the analyses surveyed by Smith. Smith considers exclusively event causation, but it may be disputed whether this is the correct conception to apply to God's case. Smith contends that considerations of agent causation are not germane to the discussion because we are not concerned with the relation between God (the agent) and His act of willing (the effect), but with the relation between His act of willing (an event) and the Big Bang (an event). But not all proponents of agent causation construe agent causation as a relation between an agent and his volitions. Some proponents of agent causation hold that an agent does not cause his volitions, but that by freely willing he brings about some intended event (Lowe 2002: 205–10). In the case at hand God brings about the Big Bang by an exercise of His active power. The expression 'God's willing that the Big Bang occur' properly describes an action, not an event. The event in this case is the Big Bang, and the cause of that event is God, who, by willing, brought about the Big Bang. Thus, it is simply wrong-headed to think of the Big Bang as caused by the event of God's willing rather than by God Himself.[24]

Thus, neither (4) nor (5) commends itself to us as more plausibly true than its contradictory. Smith recognizes these deficiencies of his argument, but he falls back to what he considers an impregnable position: '*c* is a cause of *e*' entails '*c* is not a logically sufficient condition of *e*' (Smith 1996: 176). This entailment precludes God's being the cause of the Big Bang because God's willing that the Big Bang occur is a logically sufficient condition of the Big Bang. This is because God is omnipotent, and thus necessarily His will is successful. There is no possible world in which God wills the Big Bang and yet the Big Bang fails to occur. Therefore, God cannot be the cause of the Big Bang.

This argument seems quite fanciful. If successful, it can be generalized to show that God cannot cause anything. Thus, precisely *because* He is omnipotent, God is utterly impotent—a curious inference! If being omnipotent entails inability to cause anything, then we are using 'cause' in a highly technical sense which is not incompatible with God's bringing about the Big Bang, which is, after all, the issue. Whether or not God 'causes' the Big Bang, it is still up to Him whether it occurs or not, and it occurs only in virtue of His willing that it occur. If it seems that bringing about the Big Bang does involve a causal relation, then we shall simply reject Smith's entailment principle. Only someone who is already a naturalist would be tempted to think that that principle is true. Thus, Smith's argument is either question-begging or not incompatible with God's bringing about the Big Bang.

Smith considers such a response and insists that it is the theist who begs the question, since in every other case of causation causes are not logically sufficient conditions of their effects. There is, he says, no justification for exempting God's alleged acts of causation from this principle. We need to have some independent reason for thinking that the relation between God and the Big Bang is a causal relation. Three things may be said about this response: (i) Since only God is omnipotent, it is hardly surprising that His case should be the sole exception to the principle that causes are not logically sufficient for their effects. God is so exceptional a being that He will in general not fit into our customary schemata. For example, it

is a general principle that '*S* believes *p*' is not a logically sufficient condition of '*p*.' But since God is essentially omniscient, in God's case His believing *p* is a logically sufficient condition of *p*. Should we therefore conclude that God has no beliefs? In the same way, because God is omnipotent, are we to think that His will has no effects? (ii) There are other plausible counter-examples to Smith's principle. For example, change is plausibly a cause of the existence of time, at least on a relational view of time. The occurrence of events actually brings time into existence. If there were an absolutely quiescent state, then time would not exist. But if a change occurs, time is immediately produced. Such a relation is plausibly causal; it is certainly not like the purely logical relation between, say, a two-dimensional figure's having three sides and its having three angles. Time is something altogether distinct from change, since time can go on, even most relationalists agree, even though change should cease (Shoemaker 1969). Thus, change, should it occur, would seem to cause time to exist. Yet change necessarily causes time: there is no possible world in which change is going on without time. Change is thus logically sufficient for the existence of time, but is also plausibly a cause of time's existence. (iii) The reason that the relation between God and the Big Bang—or any other event He brings about—is causal is the close resemblance between God and ourselves as agents. Doubtless our deepest intuitions about causality are rooted in our own ability to bring about effects by an intentional exertion of our power. But God is a personal agent like ourselves. The difference between Him and us is that His power is so great that He is infallible in bringing about His undertakings. Is His status as a cause now to be doubted because He is infallible? Hardly! In short I do not think that Smith's objection poses a serious obstacle to thinking that the Big Bang has a supernatural or divine cause.

All of the above objections have been considered as attempted justification of the apparently incredible position that the universe sprang into being uncaused out of nothing. But I, for one, find the premises of those objections far less perspicuous than the proposition that *whatever begins to exist has a cause.* It is far more plausible to deny one of those premises than to affirm what Hume called the 'absurd Proposition' that something might arise without a cause,[25] that the universe, in this case, should pop into existence uncaused out of nothing.

Conclusion

We can summarize our argument as follows:

7. Whatever begins to exist has a cause of its existence.
8. The universe began to exist.
9. Therefore, the universe has a cause of its existence.

Premise (7) is an intuitively grasped, metaphysical first principle. Premise (8) is supported by the inductive evidence of contemporary cosmology and enjoys greater plausibility in light of that evidence than its contradictory. An analysis of what it is to be cause of the universe reveals that

10. If the universe has a cause of its existence, then an uncaused, personal Creator of the universe exists, who *sans* the universe is beginningless, changeless, immaterial, timeless, spaceless, and enormously powerful.

From (9) and (10), it follows that

11. Therefore, an uncaused, personal Creator of the universe exists, who *sans* the universe is beginningless, changeless, immaterial, timeless, spaceless, and enormously powerful.

And this, as Thomas Aquinas laconically remarked, is what everyone means by 'God.'[26]

Notes

1 For a helpful introduction to these two competing perspectives, see Gale (1968a): 65–85.
2 See discussion in Craig (1990): 150–6.
3 This is the salient point of Grünbaum's critique of the inference to a First Cause of the origin of the universe (Grünbaum: 2000). As a B-theorist Grünbaum does not believe that the universe ever came into being, even if it had a first temporal interval. As he elsewhere writes, 'coming *into* being (or "becoming") is *not* a property of *physical* events themselves but only of human or conscious awareness of these events' (Grünbaum 1967: 153). What Grünbaum fails to see, however, is that the claim that an absolute beginning of the universe entails that the universe came into being is rooted, not in the presupposition of the so-called Spontaneity of Nothingness, but in an A-theory of time.
4 See my companion volumes *The Tensed Theory of Time: a Critical Examination,* Synthese Library (Dordrecht: Kluwer Academic Publishers, (2000a); *The Tenseless Theory of Time: a Critical Examination*, Synthese Library (Dordrecht: Kluwer Academic Publishers, 2000b).
5 For an outstanding defense of this point, see Smith (1991). See also Craig (1996a), Craig (1996b), and Craig (2000c).
6 One of the most eloquent spokesmen for this point of view has been Schlesinger (1980), 34–9, 138–9. See also Craig (1999a); Craig (1999b); and Craig (2001).
7 The most helpful here are still Broad (1938) and Dummett (1960): 497–504. See also Craig (1998a): 122–7.
8 This point needs further work, but see Prior (1968): 1–14, and Loizou (1986): 44–5. See also Craig (2003); Craig (2000d).
9 See Smith (1991) chap. 7. See also Craig (2002).
10 See remarks of Black (1962).
11 For adumbrations of this argument see Gale (1968b): 90–7 and Mellor (1981): 140. For a fuller development see Craig (forthcoming).
12 Again, this point needs to be better developed, but see Geach (1972): 306, and McGilvray, (1979): 275–99.
13 See the excellent study by Merricks (1994); see further Hinchliff, (1994); Lewis (1986); van Inwagen (1990).

14 See the very interesting recent discussions about the warrant for methodological naturalism in science, *e.g.* de Vries (1986); Plantinga, Van Till, Pun, and McMullin, (1991); Hasker (1992); Plantinga (1992); Plantinga (1993); Moreland (1994); Moreland, Meyer, and Bube (1994).

15 Smith (1993a): 135. Elsewhere he has written,

> '[This world] exists non-necessarily, improbably, and causelessly. It exists *for absolutely no reason at all* . . . The impact of this captivated realization upon me is overwhelming. I am completely stunned. I take a few dazed steps in the dark meadow, and fall among the flowers. I lie stupefied, whirling without comprehension in this world through numberless worlds other than this' (Smith 1986: 300–1).

In *Theism, Atheism, and Big Bang Cosmology,* Smith claimed that the universe came into being uncaused out of nothing at the Planck time; but he has since recanted that position under the realization that quantum cosmologists are then studying a complete fiction!

16 John Leslie asks incredulously, 'How could such nonsense have been churned out by the author of *Physical Chemistry,* a superb textbook?' (Leslie 1993: 3). For a good critique of Atkins, see Ward (1996),chap. 1.

17 See Weingard, (1982): 235–42.

18 Grünbaum (1989). For a response, see Craig (1992).

19 Grünbaum (1991). For a response, see Craig (1994a).

20 Brian Leftow puts this nicely when he writes,

> 'If God existed in time once time existed and time had a first moment, then God would have a first moment of existence: there would be a moment before which He did not exist, because there was no 'before' that moment . . . Yet even if He . . . had a first moment of existence, one could still call God's existence unlimited were it understood that He would have existed even if time did not. For as long as this is true, we cannot infer from God's having had a first moment of existence that God *came into* existence or would not have existed save if time did" (Leftow 1991: 269; cf. 201).

Senor has dubbed such a model of divine eternity 'accidental temporalism' (Senor 1993: 88). See further Craig (1996c).

21 Grünbaum (1994). For a response, see Craig (1994b).

22 See discussion and references in Craig (1998b).

23 In Smith (1993b), Smith actually arrives at this conclusion himself. He states,

> 'extant definitions of causality are incorrect since they do not cohere in the proper way with the concept of a cause of the universe . . .
>
> This entails that either there is some other (as yet undiscovered) definition of a cause that is correct or that a cause is indefinable. In the latter case, the concept of a cause would be primitive and the causal relation a simple relation known only by ostension (as is arguably the case with such relations as *being in contact with* or *being earlier than*). I know of no means of discovering or formulating a correct definition of a cause and know of no reason to think that there is such a definition. Accordingly, I think it is reasonable to conclude that the causal relation is indefinable.
>
> One way to avoid this conclusion would be to reject the assumption that the various examples of causes of the big bang . . . are genuine examples of causes . . . I would say that claims that God's creation of the big bang singularity and other

examples given . . . are not cases of possible causation are counterintuitive and are *ad hoc* attempts to retain a counterexampled theory. It is more plausible to think that a cause cannot be defined than to think that a mind's creation of a big bang singularity could not be a causal act" (Smith 1993b: 1, 24).

Smith came to think God's relation to the Big Bang is not causal because no cause is logically sufficient for its effect. But Smith does not justify why the actions of an omnipotent being would not be exceptions to this rule.

24 See Moreland (1998). I am indebted to my colleague for several interesting discussions pertinent to agency and creation.

25 David Hume to John Stewart, February, 1754, in Greig (1932), 1: 187.

26 Thomas Aquinas *Summa theologiae* 1a.2.3.

Bibliography

Atkins, P.W. (1992) *Creation Revisited*, New York: W. H. Freeman.

Black, M. (1962) Review of *The Natural Philosophy of Time, Scientific American* 206, 181–183.

Bludman, S.A. (1984) 'Thermodynamics and the End of a Closed Universe', *Nature* 308, 319–22.

Borde, A. and Vilenkin, A. (1994) 'Eternal Inflation and the Initial Singularity', *Physical Review Letters* 72, 3305–3308

Broad, C. D. (1938) *An Examination of McTaggart's Philosophy*, 2 vols., Cambridge: Cambridge University Press.

Bunge, M. (1985) *Treatise on Basic Philosophy, vol. 7: Epistemology and Methodology III: Philosophy of Science and Technology: Part I: Formal and Physical Sciences*, Dordrecht: D. Reidel.

Craig, W. L. (1990) *Divine Foreknowledge and Human Freedom: The Coherence of Theism I: Omniscience*, Studies in Intellectual History 19, Leiden: E. J. Brill.

Craig, W.L. (1992) 'The Origin and Creation of the Universe: a Reply to Adolf Grünbaum', *British Journal for the Philosophy of Science* 43:2, 233–40.

Craig, W.L. (1994a) 'Prof. Grünbaum on Creation', *Erkenntnis* 40:3, 325–41.

Craig, W.L. (1994b) 'A Response to Grünbaum on Creation and Big Bang Cosmology', *Philosophia naturalis* 31, 237–49.

Craig, W.L. (1996a) 'Tense and the New B-Theory of Language', *Philosophy* 71, 5–26.

Craig, W.L. (1996b) 'The New B-Theory's *Tu Quoque* Argument,' *Synthese* 107, 249–69.

Craig, W.L. (1996c) 'Timelessness and Creation', *Australasian Journal of Philosophy* 74:4, 646–56.

Craig, W.L. (1997) 'Hartle-Hawking Cosmology and Atheism', *Analysis* 57:4, 291–95.

Craig, W.L. (1998a) 'McTaggart's Paradox and the Problem of Temporary Intrinsics', *Analysis* 58, 122–27

Craig, W.L. (1998b) 'Divine Timelessness and Personhood', *International Journal for Philosophy of Religion* 43:2, 109–24.

Craig, W.L. (1999a) 'Tensed Time and our Differential Experience of the Past and Future', *Southern Journal of Philosophy* 37, 515–37.

Craig, W.L. (1999b) 'The Presentness of Experience', in M. Wegener (ed.) *Time, Creation, and World Order*, Aarhus, Denmark: University of Aarhus Press, 107–120.

Craig, W.L. (2000a) *The Tensed Theory of Time: a Critical Examination*, Synthese Library, Dordrecht: Kluwer Academic Publishers.
Craig, W.L. (2000b) *The Tenseless Theory of Time: a Critical Examination*, Synthese Library, Dordrecht: Kluwer Academic Publishers.
Craig, W.L. (2000c) 'On Truth Conditions of Tensed Sentence Types', *Synthese* 120, 265–270.
Craig, W.L. (2000d) 'The Extent of the Present', *International Studies in the Philosophy of Science*, 14, 165–86.
Craig, W.L. (2001) 'On Wishing It Were Now Some Other Time', *Philosophy and Phenomenological Research* 62, 159–66.
Craig, W.L. (2002) 'Divine Eternity and the Special Theory of Relativity', in G. Ganssle and D. Woodruff (eds.) *God and Time*, New York: Oxford University Press, 129–52.
Craig, W.L. (2003) 'In Defense of Presentism', in eds., Aleksander Jokic and Quentin Smith, *Time, Tense, and Reference*, Cambridge, Mass.: MIT Press, 390–408.
Craig, W.L. (forthcoming) 'Tooley on Tense and Temporal Priority', *Analysis*.
Davies, P.C.W. (1978) 'Spacetime Singularities in Cosmology and Black Hole Evaporation,' in *The Study of Time III: Proceedings of the Third Conference of the International Society for the Study of Time Alpbach – Austria*, ed. J. T. Fraser, Nathaniel Morris Lawrence; David Allen Park New York: Springer-Verlag, Berlin, 74–91.
Davies, P.C.W. (1995) 'The Birth of the Cosmos', in Jill Gready (ed.) *God, Cosmos, Nature and Creativity: The Templeton Lectures, University of Sydney 1991–1994*, Edinburgh: Scottish Academic Press, 1–27.
DeWitt, B. S. (1983) 'Quantum Gravity', *Scientific American* 249, 6, December, 104–115.
Dummett, M. (1960) 'A Defense of McTaggart's Proof of the Unreality of Time', *Philosophical Review* 69, 497–504.
Eddington, A. (1933), *The Expanding Universe*, New York: Macmillan.
Einstein, A. (1952) 'Cosmological Considerations on the General Theory of Relativity,' in A. Einstein et al. *The Principle of Relativity*, New York: Dover Publications, 177–88.
Friedmann, A. (1922), 'Über die Krümmung des Raumes', *Zeitschrift für Physik* 10, 377–86.
Gale, R.M. (1968a) 'The Static versus the Dynamic Temporal: Introduction', in R.M. Gale (ed.) *The Philosophy of Time: A Collection of Essays*, New Jersey: Humanities Press, 65–85.
Gale, R. M. (1968b) *The Language of Time*, International Library of Philosophy and Scientific Method, London: Routledge and Kegan Paul.
Gasperini, M. (1999) 'Looking Back in Time beyond the Big Bang', *Modern Physics Letters A* 14:16, 1059–1066.
Gasperini, M. (2000) 'Inflation and Initial Conditions in the Pre-Big Bang Scenario' *Physics Review D* 61, 87301–87305.
Geach, P. (1972), 'Some Problems about Time' in *Logic Matters*, Berkeley: University of California Press, 302–18.
Gott, J.R. III (1982) 'Creation of Open Universes from de Sitter space', *Nature* 295, 304–307.
Gott, J.R. III and Li-Xin Li (1998) 'Can the Universe Create Itself?' *Physical Review D* 58 (2), 23501–23544
Gott, J.R. III, Gunn, J.E., Schramm, D.N., Tinsley, B.M. (1976) 'Will the Universe Expand Forever?' *Scientific American* 234 (March), 62–79.
Greig, J. Y. T. (ed.) (1932) *The Letters of David Hume*, 2 vols., Oxford: Clarendon Press.
Gribbin, J. (1976) 'Oscillating Universe Bounces Back', *Nature* 259, 15–16.
Gribbin, J. (1986), *In Search of the Big Bang*, New York: Bantam Books.

Grünbaum, A. (1967) 'The Anisotropy of Time', in T. Gold (ed.) *The Nature of Time*, Ithaca, N.Y.: Cornell University Press, 149–86.
Grünbaum, A. (1989) 'The Pseudo-Problem of Creation in Physical Cosmology', *Philosophy of Science* 56, 373–94.
Grünbaum, A. (1991) 'Creation as a Pseudo-Explanation in Current Physical Cosmology', *Erkenntnis* 35, 233–54.
Grünbaum, A. (1994) 'Some Comments on William Craig's "Creation and Big Bang Cosmology"', *Philosophia naturalis* 31:2, 225–36.
Grünbaum, A. (2000) 'A New Critique of Theological Interpretations of Physical Cosmology', *British Journal for the Philosophy of Science*, 51:1, 1–43.
Guth, A. (1981) 'Inflationary Universe: A Possible Solution to the Horizon and Flatness Problems' *Physical Review D* 23, 247–56.
Guth, A. and Sher, M. (1983) 'The Impossibility of a Bouncing Universe', *Nature* 302, 505–06.
Hartle, J. and Hawking, S. (1983) 'Wave Function of the Universe', *Physical Review D* 28, 2960–75.
Hasker, W. (1992) 'Evolution and Alvin Plantinga', *Perspectives on Science and Christian Faith* 44, 150–62.
Hawking, S. W. (1988) *A Brief History of Time*, New York: Bantam Books.
Hawking, S.W. and Penrose, R. (1973) 'Space-Time Singularities', in S.W. Hacking and G.F.R. Ellis (eds.) *The Large-Scale Structure of Space-Time*, Cambridge: Cambridge University Press, 256–98.
Hawking, S.W. and Penrose, R. (1996) *The Nature of Space and Time*, The Isaac Newton Institute Series of Lectures, Princeton: Princeton University Press.
Hick, J. (1960) 'God as Necessary Being', *Journal of Philosophy*, 57:22/23, 725–734.
Hochberg, D., Molina-Paris, C. and Visser, M. (1999) 'Tolman Wormholes Violate the Strong Energy Condition', *Physical Review D* 59, 44011–44019.
Hoyle, F. (1948) 'A New Model for the Expanding Universe', *Monthly Notices of the Royal Astronomical Society* 108, 372–82.
Hoyle, F. (1975a) *Astronomy Today*, London: Heinemann.
Hoyle, F. (1975b) *Astronomy and Cosmology: A Modern Course*, San Francisco: W.H. Freeman.
Hubble, E. (1929), 'A Relation between Distance and Radial Velocity among Extra-galactic Nebulae', *Proceedings of the National Academy of Science* 15, 168–73.
Hume, D. (1947) *Dialogues concerning Natural Religion*, ed. with an Introduction by N. Kemp Smith, Indianapolis: Bobbs-Merrill.
Isham, C. (1988) 'Creation of the Universe as a Quantum Process' in R.J. Russell, W.R. Stoeger and G.V. Coyne (eds.) *Physics, Philosophy and Theology: a Common Quest for Understandin*g, Vatican City: Vatican Observatory, 375–408
Isham, C. (1990) 'Space, Time, and Quantum Cosmology' paper presented at the conference 'God, Time, and Modern Physics', March.
Isham, C. (1994) 'Quantum Cosmology and the Origin of the Universe', lecture presented at the conference 'Cosmos and Creation', Cambridge University, 14 July.
Jaki, S.L. (1974) *Science and Creation*, Edinburgh: Scottish Academic Press.
Kanitscheider, B. (1990) 'Does Physical Cosmology Transcend the Limits of Naturalistic Reasoning?' in P. Weingartner and G. J. W. Doen (eds.) *Studies on Mario Bunge's "Treatise'*, Amsterdam: Rodopi, 337–50.

Lamaitre, G. (1927), 'Un univers homogène de masse constante et de rayon croissant, rendant compte de la vitesse radiale des nébuleuses extragalactiques", *Annales de la Société scientifique de Bruxelles* 47, 49–59.
Leftow, B. (1991) *Time and Eternity*, Cornell Studies in Philosophy of Religion, Ithaca, N.Y.: Cornell University Press.
Leibniz, G. W. (1951), *The Monadology and Other Philosophical Writings*, trans. Robert Latta, London: Oxford University Press.
Leslie, J. (1993) 'Is It All Quite Simple?', *Times Literary Supplement*, 29 January, 3–4.
Lewis, D. (1986) 'Persons, Morality, and Tenselessness', *Philosophy and Phenomenological Research* 47, 305–9.
Lifschitz, E.M. and Khalatnikov I.M. (1963) 'Investigations in Relativist Cosmology', *Advances in Physics* 12, 185–249.
Linde, A.D. (1983) 'Chaotic Inflation', *Physics Letters* 129, 177–81.
Linde, A.D. (1984) 'The Inflationary Universe', *Reports on Progress in Physics* 47, 925–86.
Linde, A.D. (2002) 'Cyclic Universe Runs into Criticism', *Physics World*, June, 8.
Linde, A., Linde, D. and Mezhlumian, A. (1994) 'From the Big Bang Theory to the Theory of a Stationary Universe' *Physical Review D* 49, 1783–1826.
Loizou, A. (1986) *The Reality of Time*, Brookfield, Vt.: Gower.
Lowe, J. (2002) *A Survey of Metaphysics*, Oxford: Oxford University Press.
Malcolm, N. (1958) *Ludwig Wittgenstein: A Memoir*, London: Oxford University Press.
McGilvray, J. (1979) 'A Defense of Physical Becoming', *Erkenntnis* 14: 275–99.
Mellor, D.H. (1981) *Real Time*, Cambridge: Cambridge University Press.
Merricks, T. (1994) 'Endurance and Indiscernibility', *Journal of Philosophy* 91, 165–84.
Hinchliff, M. (1994) 'The Puzzle of Change', paper presented at the Pacific Division meeting of the American Philosophical Association, April 2.
Moreland, J.P. (1994) 'Theistic Science and Methodological Naturalism', in J.P. Moreland (ed.) *The Creation Hypothesis*, Downer's Grove, Ill.: Inter-Varsity Press, 41–66.
Moreland, J. P. (1998) 'Libertarian Agency and the Craig/Grünbaum Debate about Theistic Explanation of the Initial Singularity', *American Catholic Philosophical Quarterly* 81, 539–54.
Moreland, J. P., Meyer, S. C., and Bube, R. H. (1994) 'Conceptual Problems and the Scientific Status of Creation Science: a Discussion', *Perspectives on Science and Christian Faith* 46, 2–25.
Naber, G.L. (1988), *Spacetime and Singularities: an Introduction*, Cambridge: Cambridge University Press.
Novikov, I.D. and Zel'dovich, Y. B. (1973) 'Physical Processes near Cosmological Singularities', *Annual Review of Astronomy and Astrophysics* 11, 387–412.
Parfit, D. (1998) 'Why Anything? Why This?' *London Review of Book*, 20:2, 24–27.
Penrose, R. (1965) 'Gravitational Collapse and Space-Time Singularities', *Physical Review Letters* 14, 57–59.
Penrose, R. (1981) 'Time-Asymmetry and Quantum Gravity', in C.G. Isham, R. Penrose, and D.W. Sciama (eds.) *Quantum Gravity 2*, Oxford: Clarendon Press, 245–72.
Penrose, R. (1982) 'Some Remarks on Gravity and Quantum Mechanics' in M.J. Duff and C.J. Isham (eds.) *Quantum Structure of Space and Time*, Cambridge: Cambridge University Press, 3–10.
Penrose, R. (1997) *The Large, the Small, and the Human*, Cambridge: Cambridge University Press.

Plantinga, A. (1992) 'On Rejecting The Theory of Common Ancestry: A Reply to Hasker', *Perspectives on Science and Christian Faith* 44, 258–63.
Plantinga, A. (1993) 'Methodological Naturalism', paper presented at the symposium 'Knowing God, Christ, and Nature in the Post-Positivistic Era', University of Notre Dame, April 14–17.
Plantinga, A., Van Till, H.J., Pun, P., and McMullin, E. (1991) 'Symposium: Evolution and the Bible', *Christian Scholar's Review* 21, 8–109.
Post, J. (1991), *Metaphysics: a Contemporary Introduction*, New York: Paragon House.
Prior, A. N. (1968) 'Changes in Events and Changes in Things', in *Papers on Time and Tense*, Oxford: Clarendon Press, 1–14.
Rees, M. (1997) *Before the Beginning*, with a Foreword by S. Hawking, Reading, Mass.: Addison-Wesley.
Reeves, H., Audouze, J., Fowler, W.A. and Schramm, D.N. (1973) 'On the Origin of Light Elements', *Astrophysical Journal* 179, 909–930.
Russell, B. and Copleston, F.C. (1964) 'The Existence of God', in J. Hick (ed.) *The Existence of God*, New York: Macmillan, 167–191.
Schlesinger, G. (1980) *Aspects of Time*, Indianapolis: Hackett, 1980.
Senor, T. D. (1993) 'Divine Temporality and Creation *ex nihilo*', *Faith and Philosophy* 10, 86–92.
Shoemaker, S. (1969) 'Time without Change', *Journal of Philosophy* 66, 363–81.
Silk, J. (1989) *The Big Bang*, 2d ed., San Francisco: W.H. Freeman.
Smith, Q. (1986) *The Felt Meanings of the World*, Lafayette, Ind.: Purdue University Press.
Smith, Q. (1991) *Language and Time*, New York: Oxford University Press.
Smith, Q. (1993a) 'The Uncaused Beginning of the Universe', in W.L. Craig and Q. Smith (eds.) *Theism, Atheism, and Big Bang Cosmology*, Oxford: Clarendon Press, 108–40.
Smith, Q. (1993b) 'The Concept of a Cause of the Universe', *Canadian Journal of Philosophy* 23, 1–24.
Smith, Q. (1996) 'Causation and the Logical Impossibility of a Divine Cause', *Philosophical Topics* 24, 169–91.
Suarez, F. (2002) *On Creation, Conservation, and Concurrence: "Metaphysical Disputations 20, 21 and 22"*, trans. with an Introduction and Notes by A.J. Freddoso, South Bend, Ind.: St. Augustine's Press.
Swinburne, R. (1991) *The Existence of God*, rev. ed. Oxford: Clarendon Press.
Tryon, E. (1973) 'Is the Universe a vacuum Fluctuation?' *Nature* 246, 396–97
Van Inwagen, P. (1990) 'Four Dimensional Objects', *Noûs* 24, 245–55.
Vilenkin, A. (1982) 'Creation of the Universe from Nothing', *Physical Letters B* 117, 25–28.
Vilenkin, A. (1983) 'Birth of Inflationary Universes', *Physical Review D* 27, 2848–55.
Vries, P. de (1986) 'Naturalism in the Natural Sciences: A Christian Perspective', *Christian Scholar's Review*, 388–96.
Ward, K. (1996), *God, Chance, and Necessity*, Oxford: One World.
Weingard, R. (1982), 'Do Virtual Particles Exist?', in P. Asquith and T. Nichols (eds.) *Proceedings of the Philosophy of Science Association*, 2 vols., East Lansing, Mich.: Philosophy of Science Association, 1, 235–42.
Wheeler, J.A. (1980) 'Beyond the Hole', in H. Woolf (ed.) *Some Strangeness in the Proportion: A Centennial Symposium to Celebrate the Achievements of Albert Einstein*, Reading, Mass.: Addison-Wesley, 341–375.
Yates, J.C. (1990) *The Timelessness of God*, Lanham, MD.: University Press of America.

Permissions

Thanks to the publishers listed below for their permission to reprint previously published material for this volume of the anthology.

Permission has been granted by John Wiley and Sons (www.wiley.com) to reprint the following excerpted material:

William Lane Craig and James Sinclair, "science" section excerpted from "The Kalam Cosmological Argument," in *The Blackwell Companion to Natural Theology*, 125–82, ed. William Lane Craig and J. P. Moreland. Oxford: Blackwell, 2009.

Special thanks to Dr. J. R. Gott and Dr. Maurizio Gasperini for their permission to use their graphics (Figure 1.3, Gott; 15 and 16, Gasperini) found in the following essay: William Lane Craig and James Sinclair, "science" section excerpted from "The Kalam Cosmological Argument," in *The Blackwell Companion to Natural Theology*, ed. William Lane Craig and J. P. Moreland. Oxford: Blackwell, 2009. Thanks to James Sinclair for permission to use the other nonpublic domain graphics and figures throughout the rest of this excerpted piece.

Permission has been granted by Oxford University Press to reprint the following article:

J. B. Pitts "Why the Big Bang Singularity Does Not Help the Kalam Cosmological Argument for Theism." *British Journal for the Philosophy of Science* 59, no. 4 (October 2008): 675–708.

Permission has been granted by Macmillan and Palgrave Macmillan to reprint the following material:

William Lane Craig and James Sinclair, "On Non-Singular Spacetimes and the Beginning of the Universe." With James Sinclair. In *Scientific Approaches to the Philosophy of Religion*, 95–142. Ed. Yujin Nagasawa. Palgrave Frontiers in Philosophy of Religion. London: Macmillan, 2012. Reproduced with permission of Palgrave Macmillan.

Permission for the reprinting of the excerpted article "Beyond the Big Bang" has been granted by Taylor & Francis, Ltd. This excerpt was taken from the following essay:

William Lane Craig. "Naturalism and Cosmology," in *Analytic Philosophy without Naturalism*, 97–133, ed. A. Corradini, S. Galvan, and E. J. Lowe, *Routledge Studies in Contemporary Philosophy*. New York: Routledge, 2006.

The same figure by J. R. Gott mentioned above is included in William Lane Craig's excerpted essay, "Beyond the Big Bang" (Figure 11.1). This piece was taken from "Naturalism and Cosmology," in *Analytic Philosophy without Naturalism*, 97–133,

eds. A. Corradini, S. Galvan, and E. J. Lowe, Routledge Studies in Contemporary Philosophy. New York: Routledge, 2006.

Permission has been granted by the American Physical Society to publish the following article from its journal, *Reviews of Modern Physics*:

Fred Adams and Gregory Laughlin, "A Dying Universe: The Long-Term Fate and Evolution of Astrophysical Objects," *Reviews of Modern Physics* 69, no. 2 (April 1997): 337–72.

Thanks to Fred Adams and Gregory Laughlin for permission to use the graphics from their article in this anthology.

Thanks to the editor of the journal *Classical and Quantum Gravity*, who gave special permission to publish the following article:

Aron C. Wall, "The Generalized Second Law implies a Quantum Singularity Theorem," arXiv: 1010.5513v3 [gr-qc] August 12, 2013. © IOP Publishing. Reproduced with permission. All rights reserved. Published July 12, 2013 • 2013 IOP Publishing Ltd Classical and Quantum Gravity, Volume 30, Number 16, 165003

The author has made subsequent corrections to the published version. This version is available online as arXiv:1010.5513v5 and is being used in this anthology.

In addition, we are grateful to Aron Wall, who not only granted personal permission to use his article for this anthology, but he provided a revised and updated version of his article to use in the anthology. (Wall blogs on physics and theology at "Undivided Looking"—http://www.wall.org/~aron/blog/.)

Thanks to the executive editor Steven Wheeler of the journal *Inference: International Review of Science* (www.inference-review.com) for permission to include the following article, originally published in *Inference*:

Alexander Vilenkin, "The Beginning of the Universe." *Inference: International Review of Science* (2015). http://inference-review.com/article/the-beginning-of-the-universe.

Permission has been granted by Springer to reprint the following articles from *Open Systems & Information Dynamics*:

Gábor Kutrovátz, "Heat Death in Ancient and Modern Thermodynamics," *Open Systems and Information Dynamics* 8, no. 4 (December 2001): 349–59. Doi: 10.1023/A:1013901920999. Print ISSN 1230–1612. Online ISSN 1573–1324. Journal number: 11080. With permission of Springer.

Special thanks to Gábor Kutrovátz for permission to use his graphics from his aforementioned article.

Milan M. Cirkovic, "Entropy and Eschatology: A Comment on Kutrovátz's Paper 'Heat Death in Ancient and Modern Thermodynamics,'" *Open Systems and Information Dynamics* 9, no. 3 (September 2002): 291–99. Doi: 10.1023/A:1019716815925. Print ISSN 1230–1612. Online ISSN 1573–1324. Journal number: 11080. With permission of Springer.

Permission to reprint the following article from the *American Catholic Philosophical Quarterly* has been granted by the Philosophy Documentation Center of Charlottesville, Virginia (www.pdcnet.org):

J. P. Moreland, "Libertarian Agency and the Craig/Grünbaum Debate about Theistic Explanation of the Initial Singularity," *American Catholic Philosophical Quarterly* 71, no. 4 (Autumn 1997): 539–54. DOI: 10.5840/acpq199771419

The Permissions Company, Inc., has granted permission for reprinting the following essay:

Quentin Smith, "Causation and the Logical Impossibility of a Divine Cause" from *Philosophical Topics* 21, no. 1 (Spring 1996): 169–91. Copyright © 1996 by the Trustees of the University of Arkansas. Reprinted with the permission of The Permissions Company, Inc., on behalf of the University of Arkansas Press, www.uapress.com.

Index

acausal fine-tuning, problem of 30, 44
accidental temporalism 347 n.20
active power 301, 310 n.6
Adams, F. 8
Adams, M. 301
Advaita Vedanta school of Hinduism 2
agency/agent causation 9, 299–309
Aguirre, A. 44, 128, 131–3, 143 n.14
al-Ghāzālī 2, 4, 110
Ambjorn, J. 138, 139
analytic continuation 65
Anaxagoras 244
Anaximandros (i.e., Anaximander) 244, 250 n.1
Anselm 2
anti-de Sitter
  causality in 281
  and de Sitter space 275
Aquinas, T. 2–3, 83, 84, 86, 346
Aristotle 2, 9, 82
  physics of 232–4
Aronson, J. 315
Aronszajn, M. 330 n.1
arrow of time (AOT) 128
Ashtekar, A. 92, 97
asymptotically static universe 33–7, 124–6
asymptotic giant branch (AGB) 166, 167
asymptotic past triviality (APT) 48, 49, 52
A-theory 334–5, 336, 346 n.3
  arguments for and against 335
Atkins, P.W. 338

baby universe 272–5, 281, 291 n.13
Bach-Weyl formalism 122
Bach–Weyl theory of gravity 81, 84–5, 102 n.1, 115
background radiation fields 207–212
Banks, T. 43, 72 n.35, 73 n.35, 74 n.48, 127, 143 n.22
Barrow, J. 74 n.53
baryon number violations 184, 185, 196–8
Baum-Frampton approach 38–40, 43
Bekenstein bound 226
Bekenstein-Hawking formula 246
Big Bang cosmological model 6, 16–17, 33, 34, 35, 38, 45, 68, 70 n.5, 75 n.60, 80–1, 89, 95, 96–8, 100–1, 111, 160, 164, 204, 205
Big Crunch 7, 31, 43, 57, 60, 63, 68, 73 n.35, 74 n.49, 114, 202, 247, 291 n.13
Big Rip 38, 41, 42, 45, 73 n.34
Bishop, J. 306, 308, 311 n.26
BKL chaos 30–1, 44, 45, 55, 57, 68, 124, 131
Black Crunch 73 n.35, 127, 144 n.22
black hole 208, 253, 257
  accretion time 178–9
  baby universes and 271–275
  continued formation and decay of 216–18
  decay 130
    and Hawking radiation 200–1
  entropy 58–60
  era of 224
  evaporation 209–10, 217, 273–5
  Hawking radiation and 258
  horizons 259
Blackwell Companion to Natural Theology (Craig and Sinclair) 8
Bojowald, M. 36, 37, 57–61, 74 n.50, 99, 125, 127, 128, 133, 134
Bonaventure 2, 83, 86
Borde, A. 28, 120, 333
Borde-Guth-Vilenkin Theorem (BGV) 6, 8, 34, 37, 39, 44, 56, 62, 67, 68, 71 n.18, 121, 122, 124, 126, 131, 143 n.14, 152–3, 155, 157
  exceptions to 30
Bousso, R. 43, 62
branes 53, 54

Brief History of Time (Hawking) 6
brown dwarfs 161, 169, 170
collisions, star formation through 174–8
B-theory 335–6
arguments for and against 336
Bunge, M. 337

The Cambridge Companion to Atheism (Smith) 1, 141 n.1
cardinality and infinity sizes 86–7
Cassidy 70 n.7
Castañeda, H.-N. 315–16
categorical ability 301
Cauchy horizon 20, 21, 253, 271–3
Cauchy surface 101, 132, 252, 261, 267, 270, 272
Causal Dynamical Triangulations approach 139
causal horizons 38, 42–3, 259, 267, 269, 272, 284, 285, 287
causation
analogical and literal descriptions of 322–3
divine relation formulation to Big Bang 323
cosmological and teleological arguments for God's nonexistence and 328–30
and logically sufficient conditions 317–20
Sosa' theory 321
objections to God's no role in 323–8
originating divine cause notion and 312–13
cause, Hume's definition of 313–15
c-boundaries 32, 71 n.20
central pressure 165, 191–2
chaotic inflation 26–7, 71 n.14
child universe 215–16
Chisholm, R. 301
Chronology Protection Conjecture (CPC) 22–3
Cicero 250 n.1
Cirkovic, M.M. 9
Clarke, C.J.S. 101
Clarke, S. 92
Clausius, R. 233
closed timelike curves 20–3, 67
closed timelike curves (CTC) 281
closed universe 240
future expansion of 202
thermodynamic functions in 239
coarse-grained GLS 260, 277, 287
Collins, F. 80, 100
connected baby universe 273
conventional stellar evolution end 160–1
continued star formation in galaxy and 167–8
fate of earth and sun and 166–7
final mass function 168–70
main-sequence stars' lifetimes and 161–3
metallicity effects 163–6
Copan, P. 10 n.4, 115
Copan–Craig criterion 85
Copernican Time Principle 226
cosmic forgetfulness 133
cosmic strings 71 n.12
cosmological argument, versions of 2–7
cosmological wave function 62
Coule, D.H. 22, 70 n.8
counterfactual definitions, of causation 316–17
Craig, W.L. 1, 4, 7, 8, 9, 90, 115, 153, 299–300, 303, 309, 311 n.26, 330, 346 n.5
creation doctrine and warrant 82–6
creation *ex nihilo* 62–7, 82, 139, 142 n.10, 144 n.28
critical density parameter 73 n.43
Cyclic Conformal Cosmogony 122, 129
cyclic universe 37–43, 153
eternal 127–31

Dabrowski, M. 74 n.53
Damour, T. 71 n.19
dark energy 15, 32, 38, 41, 42, 48, 53, 59, 61, 72 n.29, 74 n.53, 128–9, 247
dark era 224
dark matter 39, 59, 172, 209, 212, 217, 223, 225
halo dark matter annihilation and capture 179–82
fate of planets during galactic death 182–3

Davies, P.C.W 6, 17, 333
Dawkins, R. 153
deductive argument 5
deflation 38–43
"Deflation at Turnaround for Oscillatory Cosmology" (Baum and Frampton) 41
Degenerate Era 171, 224
degenerate stellar objects long-term fate 183
  Hawking radiation and black holes decay 200–1
  higher order proton decay 196–200
  neutron stars powered by proton decay 194–6
  proton decay 184–6
    in planets 201
  white dwarf
    chemical evolution 188–90
    evolution final phase 190–4
    powered by proton decay 186–8
Democritus 244
Dennett, D. 153
density fluctuations and flat and open universe expansion 202–4
density perturbations 152, 179, 186, 203, 204, 206, 207, 216–18, 223, 226, 248
dependent baby universes 272, 274–5
De Rerum Natura (Lucretius) 244–5
Descartes, R. 2
de Sitter horizons 259, 275
de Sitter model 32, 34
de Sitter space 43, 44, 67, 130–2, 275, 291 n.19
  and anti-de Sitter space 275
DeWitt, B.S. 62
Disputationes metaphysicae (Suarez) 343
divine agent 303
divine personal explanation 304
domain walls 71 n.12
Doppler effect 16
Ducasse, C.J. 315
Ducasse's singularist definition of cause 315
dying universe 159–60, 219–26
  conventional stellar evolution end and 160–1
    continued star formation in galaxy and 167–8
    fate of earth and sun and 166–7
    final mass function 168–70
    main-sequence stars' lifetimes and 161–3
    metallicity effects 163–6
  degenerate stellar objects long-term fate and 183
    Hawking radiation and black holes decay 200–1
    higher order proton decay 196–200
    neutron stars powered by proton decay 194–6
    proton decay 184–6, 201
    white dwarf evolution final phase 190–4
    white dwarfs chemical evolution 188–90
    white dwarfs powered by proton decay 186–8
  energy and entropy production speculations, in far future
    black holes continued formation and decay 216–18
    particle annihilation in open universe and 218–19
    positronium formation and decay 219
  entropy and heat death and 225–6
  experimental and theoretical implications 224–5
  future universe ears and 224
  galaxy death and 170–1
    black hole accretion time 178–9
    dynamical relaxation of galaxy 171–3
    gravitational radiation and decay of orbits 173–4
    star formation through brown dwarf collisions 174–8
  halo dark matter annihilation and capture 179–82
    fate of planets during galactic death 182–3

long-term evolution of universe and 201–2
background radiation fields 207–212
closed universe future expansion 202
density fluctuations and flat and open universe expansion 202–4
inflation and universe future 204–7
vacuum energy density possible effects 212–16
Dyson, F.J. 249, 250

Earman, J. 94, 100, 101, 141 n.5
Eddington, A. 18, 60
Einstein, A. 12–13, 18, 26, 93, 125, 143 n.13, 150
Einstein-de Sitter universe 248
Einstein static state (ESS) 35, 36, 52, 53, 66, 68, 72 n.26, 124–6, 143 nn.20–1
Ekpyrotic/cyclic model 53–6, 74 nn.46–7, 142 n.11
Ellis, G.F.R. 29, 31, 34, 72 nn.20, 24, 74 n.48, 123–5, 143 n.17, 152–3
Emergent model 34–6, 47, 48, 52, 68, 124–6, 140, 143 n.17
emergent universe 34–7
from metastable loop-quantum-gravity state 36
Empedocles 244
energy and entropy production speculations, in far future
black holes continued formation and decay 216–18
particle annihilation in open universe and 218–19
positronium formation and decay 219
Enquiry (Hume) 1
entropy 38–40, 42, 45, 57–61, 68, 73 n.35, 74 nn.49, 51, 127–31, 144 nn.22–3, 153, 158 n.12, 241, 248, 256
black-hole 58–60
and eschatology 243–5
modern cosmology and errors in Kutrovátz's account 246–9
physical eschatology and open systems 249–50
fluctuations 275, 282–3, 291 n.19
generalized 258, 260–71
gravitational 246
heat death and 225–6
horizon 258
increase in 257, 259
Planck 127
ergodic property 43, 73 n.35, 127, 143–4 n.22
eternal inflation 23–9, 151–2
Euclidean metric 65, 75 n.57
Euclidean quantum gravity 138
event causation 300
expansion parameter 157
extropy 235, 237, 240, 241
in open universes, and heat death definition and 248–9

Fair, D. 315
false vacuum decay 152
Feynman, R. 64, 138
Feynman diagram 184
final mass function 168–70, 176, 222
fine-grained GLS 259–60, 270, 271, 276, 278, 283
First Antinomy 5
First Movent 234
Fischler, W. 73 n.35, 127, 144 n.22
"Five Ways" (Aquinas) 2–3
Flew, A. 80
Fourth Lateran Council 85–6, 101
Frampton, P. 38–40, 42–3, 72 n.30
Freivogel, B. 43
Friedman equation, modified 40–1
Friedmann, A. 6, 13–15, 150
Friedmann-Lemaître model 15–17, 33
Friedmann-Robertson-Walker (FRW) cosmology 276–7
Friedman universe 32
frozen stars 166
future horizon 274, 277, 281, 285

galaxy death 170–1
black hole accretion time and 178–9
dynamical relaxation of galaxy and 171–3

gravitational radiation and decay of orbits and 173–4
star formation through brown dwarf collisions and 174–8
Gasperini, M. 46–8, 51, 52, 56, 143 n.19
generalized entropy monotonicity properties 260–9
generalized second law (GSL) 252–4
applications
Big bangs and beginnings 276–7
black holes and babies 271–5
time machines 281
warp drives and negative mass objects 278–81
coarse-grained 260, 277, 287
entropy fluctuations and 275, 282–3, 291 n.19
fine-grained 259–60, 270, 271, 276, 278, 283
generalized thermodynamics theorems 260–71
quantum geometries and 283–6
second law of thermodynamics and 254–60
generalized thermodynamics theorems
generalized entropy monotonicity properties and 260–9
quantum trapped surfaces and 269–71
general theory of relativity (GTR) 6, 12, 24, 80, 87, 92, 94, 96–8, 125, 126, 131, 141 n.6, 141–2 n.8, 334
evidence for 18–19
exceptionalism 91
generic energy condition 20
geodesic 16–18, 20, 28, 39, 52, 53, 55, 69 n.4, 152
congruence 156, 157, 158 n.20
incomplete 151
Geroch, R. 110
Gibbs entropy 354
Gilkey, L. 83
global hyperbolicity 270, 271, 274, 276, 277, 279, 281, 285–6, 290 n.12
God, proof of 153–4
God-of-the-gaps argument 98–9
Goetz, S.C. 301
Gott III, J.R. 20, 70 nn.7–9, 139, 145 n.31, 333, 334
Gott-Li model 20–3, 65
Gould, S.J. 93
The Grand Design (Hawking) 6
Grand Unified Theories (GUTs) 184–5
Gratton, S. 44, 131–3
gravitating systems thermodynamical capacities 246–7
gravitational radiation and decay of orbits 173–4
Greene, B. 24, 25
Gribbin, J. 338
Grünbaum, A. 9, 299, 300, 303, 304, 306–9, 328, 339, 340, 346 n.3
Guth, A. 23, 28, 44, 71 n.17, 120, 123, 151, 158 n.6

Hackett, S.C. 1–2, 4, 10 nn.4, 6
Hamiltonian constraint 99
Hartle, J. 20, 137
Hawking, S. 6, 18, 20, 21, 66, 70 n.7, 135, 137–9, 142 n.10
Hawking-Penrose theorems, exceptions to 19–20
asymptotically static space time 33–7
closed timelike curves 20–3
cyclic universe 37–43
eternal inflation 23–9
infinite contraction 29–33
quantum gravity 45–67
time deconstruction 44
Hawking radiation and black holes decay 200–1
heat death 225
alternative answer to 243–5
ancient solution to 233–4
definition, and extropy in open universes and 248–9
and entropy 225–6
modern solution to 235
beyond model 240–1
extropy 235
model universe 235–40
problem of 232–3
Heisenberg's uncertainty principle 65, 67, 139
Henneaux, M. 71 n.19

Hertzsprung-Russell diagram
complete evolution of sun in 195
for low-mass stars 162
H-H formalism 138
Hilbert's Hotel 86
Hill, C. 330 n.1
Hiscock, W.A. 22
history and future of universe, important events in 220–1
horizon problem 24
horizon thermodynamics 253
Hoyle, F. 17
Hubble, E. 6, 16
Hubble parameter 13, 40, 41, 157, 202, 205, 218
Hubble radius 41–2
Hubble time 70 n.5
Hulse, R.A. 70
Hume, D. 1, 9, 313, 345
Huxley–Draper–White thesis 98
Hyppolites 250 n.1

ibn Sīna 2
inductive argument 5
ineffability theory 324–5
infinite contraction 29–33
inflation, *see also individual entries*
restarting 275
and universe future 204–7

Jurkiewicz, J. 138

The Kalām Cosmological Argument (Craig) 1
Kanitscheider, B. 337, 339
Kiefer, C. 135
Kim, J. 316
Kitab al-Iqtisad (al-Ghazali) 110
Krauss, L. 153
Kuhn, T. 93
Kutrovátz, G. 9

Lakatos, I. 93
Laughlin, G. 8
Leftow, B. 316, 347 n.20
Leibniz, G.W. 2, 3, 92–3, 101
Leibnizian cosmological argument 3
Leinbizian intuition 93, 94
LeMaître, G. 6, 13
Leslie, J. 347 n.16
Lewis, D. 316, 317, 342, 343
libertarian agency 9, 299–309
Linde, A. 26, 62, 71 n.18, 75 n.55
Li-Xin Li 20, 70 nn.7–9, 139, 145 n.31, 333, 334
Locke 2, 82
Loll, R. 138, 139
long-term evolution of universe and 201–2
background radiation fields and 207–12
closed universe future expansion and 202
density fluctuations and flat and open universe expansion and 202–4
inflation and universe future and 204–7
vacuum energy density possible effects and 212–16
loop quantum cosmology 134–5
loop quantum-gravity (LQG) 36–7, 56–62, 68, 72 n.27, 74 n.54, 97, 125
Lorentzian metric 65, 69
Lucretius 244

McGrew, T. 141 n.8
McInnis, B. 142 n.12
Mackie, J. 317
McMullin, E. 86
magnetic monopoles 71 n.12, 204
Maimonides 2, 82, 83
M dwarfs 161–3
metastability 36–7
metric conventionalism 141 n.7
Milne, E. 153
Milne phase 48, 52, 73 n.43
Misner, C. 31, 62, 137
Mixmaster universe 31, 55
modern cosmology, and errors in Kutrovátz's account 246
extropy in open universes and heat death definition and 248–9
gravitating systems thermodynamical capacities 246–7
gravitational field as major entropy source and 246
open vs. closed universes and cosmological constant and 247–8

modern cosmology and creation 87–90
Monton, B. 110–11
Moreland, J.P. 9, 348 n.24
mosaic model 38, 72 n.28
motion 3
Mulryne, D. 125

Nagel, T. 306–7
naturalism 308, 311 n.26
  objections of 337–45
The Nature of Space and Time (Hawking and Penrose) 66, 138
negative pressure 67
neutrino gas 241
neutron decay, Feynman diagram for 184
neutron stars 170
  powered by proton decay 194–6
Newton, I. 18, 92
"no-boundary" model 6, 64–5, 121, 137
nomological relatedness 314–15
non-singular space times and beginning of universe 110–11
  asymptotically static universe and 124–6
  atemporalist position 117–19
  eternally cyclic universe and 127–31
  expanding universe and 120–3
  infinitely contracting universe and 123–4
  singularity argument and 111–16
  time's arrow reversal and 131–7
  universe's beginning and 137–40
null convergence condition 150–1, 158 n.7

O'Connor, T. 309–10 n.6
open universe
  vs. closed universes, and cosmological constant and 247–8
  density fluctuations and expansion of 202–4
  extropy, and heat death definition and 248–9
  particle annihilation in 218–19
  thermodynamic functions in 238
orbital decay 166
originating divine cause, notion of 312–13
  counterfactual definitions and 316–17
  Ducasse's singularist definition of cause and 315
  Hume's definition of cause and 313–15
  transference definition of cause and 315–16

Page, D. 89
Pagels, H. 66
Parikh, M. 132, 133
particle annihilation, in open universe 218–19
Penrose, R. 6, 18, 59, 60, 66, 74 n.49, 122, 129–31, 136, 150
Penrose diagram 32–3
"phantom bounce" cosmogonies 68, 73 n.31, 143 n.21
phantom energy, *see* dark energy
Philoponus, J. 2, 4
physical eschatology and open systems 249–50
Pitts, J.B. 8, 75 n.61, 111–16, 120–2, 126, 140, 141 nn.2–6, 141–2 n.8, 144 n.25
Planck entropy 127
planets' fate, during galactic death 182–3
Plantinga, A. 1, 82
Plato 2
Plutarch 250 n.1
Polchinski, J. 62
Polkinghorne, J. 83
Pope John Paul II 101
positive energy theorem 280–1
positronium formation and decay 219
Post, J. 339
potency-act distinction, of Aristotle 3
potential infinite 5
pre-Big Bang scenario 29, 31, 44, 73 n.40, 74 n.44, 117, 119, 120, 121, 124, 126, 134, 136, 137, 253
pre-Big Bang inflation (PBBI) 46–53, 143 n.19

Principle of Sufficient Reason 3
proton decay 184–6
  Feynman diagram for 184
  higher order 196–200
  in planets 201
pycnonuclear reactions 188, 189, 200

quantum creation 154–5, 158 n.18
quantum fluctuation 36
quantum geometries 283–6
quantum gravity 6, 45, 128
  loop 36–7, 56–62, 68, 72 n.27, 74 n.54, 97, 125
  semi-classical creation *ex nihilo* models 62–7
  and singularities 97–8
  string models 46–56
quantum Lorentzian manifold 284, 285
quantum trapped surfaces 269–71
quantum tunneling 63–4
quantum vacuum 338–9
quarks 184, 185, 196, 197, 199
quasi-Milne phase 48, 53
Quinn, P. 83
radiation-dominated era 224
Raychaudhuri equation 252
Rayleigh–Jeans law, for blackbody radiation 93–4
Red Giants 163
Reformed epistemology 82, 100
reheating process 26
repulsive gravitational force 67
repulsive gravity 26
Rescher, N. 307
The Resurrection of Theism (Hackett) 1
Rindler horizons 259
Rindler vacuum 22
Robertson–Walker metric 89, 92, 94
Russell, B. 3

St. Augustine 144 n.28
Sandage, A. 89
scalar-tensor theory 84
Schlegel, R. 7
Schleiermacher, F. 83
Schrödinger, E. 132
Schücking, E. 80
Scientific American 47, 48, 50
scientific confirmation 12–16
Scotus, John Duns 2
Searle, J. 306
second law of thermodynamics 153
  Generalized Second Law 257–60
  Ordinary Second Law (OSL) 254–7
self-contradiction, in realm of being 1–2
semi-classical creation *ex nihilo* models 62–7
semi-classical quantum gravity model 68
Senor, T.D. 347 n.20
Shapiro delay 280–1
shrinking horizon 213
Sinclair, J. 8, 58–60
singularity argument 111–13, *see also individual entries*
  assessment 113
  of premises 113–16
singularity theorem 6, 9, 17, 18, 150–1, 252, 271
  exceptions to 19–67
Slipher, V. 16
Smith, Q. 1, 7, 9, 81, 135–6, 141 n.1, 144 n.25, 311 n.26, 337, 339, 342–5, 346 n.5, 347 n.15, 347–8 n.23
Smolin, L. 139
smooth particle hydrodynamics 176
Sober, E. 93
Sosa, E. 321, 329
space-time model, of four-dimensional universe 6
spallation interactions 188, 189–90
spatiotemporal continguity 314
Spengler, O. 244
sphalerons 198
Spinoza, B. 2
star formation through brown dwarf collisions 174
  collision cross sections 175–6
  collision time scales 174–5
  numerical simulations 176–8
Steinhardt, P. 47, 53–6, 74 n.45, 142 n.11, 153
stellar lifetimes vs metallicity 163–5
stellar masses vs metallicity 165–6
stelliferous era 160, 224
Stenger, V. 153

Strawson, G. 315
String Landscape 28, 62, 68
string models 46
  Ekpyrotic/cyclic model 53–6
  pre-Big Bang inflation (PBBI) 46–53
string models quantum gravity 46–56
string perturbative vacuum (SPV) 48–53, 66, 68
string vacuum 74 n.44
Stromateis 250 n.1
Suarez, F. 343
sublunary stage 233–4
Summa Theologica (Aquinas) 3
superlunary stage 234
supernaturalist alternative theory 336–7
supertask 71 n.21
Susskind, L. 62
Swinburne, R. 310 n.11

Takahashi, T. 42, 43
Tavakol, R. 128
Taylor, Joseph H., Jr. 70 n.6
temporal priority 313–14
theism 80–2
  Big Bang cosmology not congenial to faith and 100–1
  cardinality and infinity sizes and 86–7
  creation doctrine and warrant and 82–6
  fluctuating and inaccessible warrant and 99–100
  general theory of relativity and 96
  Leibniz and 92–3
  modern cosmology and creation and 87–90
  quantum gravity for singularity resolving and 97–8
  stellar collapse and 94–6
  tolerance towards singularities and 90–2
  vicious God-of-the-gaps argument and 98–9
Theism, Atheism, and Big Bang Cosmology (Craig and Smith) 7
thermodynamics, second law of 7
Thomistic cosmological argument 2
't Hooft, Gerard 122
Thorne, K. 21, 70 n.8, 94
Thucydides 244
time's arrow reversal 131–7
time deconstruction 44
time machines 281
Time Without End (Dyson) 159
Tod, P. 129
Tolman model 57, 68
Tooley, M. 331 n.3
transcendent baby universes 272
transference definition, of cause 315–16
"tunneling from nothing" approach 20, 51, 63–5, 121, 137
Turok, N. 47, 53, 55, 142 n.11, 153

"The Unique Universe" (Smolin) 119
unitary evolution 255–6, 268

Vaas, R. 37, 52, 73 n.44, 74 nn.44, 47, 126
vacuum energy density, possible effects of 212
  future inflationary epochs and 212–13
  tunneling processes and 213–16
vacuum energy density possible effects 212–16
Veneziano, G. 46–7, 51, 52, 143 n.19
viable baby universes 272–4
Vilenkin, A. 8, 20, 28–9, 38, 62–3, 66, 70 n.10, 75 n.57, 120, 127, 142 n.11, 333
von Neumann entropy 254, 255, 258

Wald, R. 16, 18, 29, 71 n.18, 97
Wall, A.C. 9
Ward, K. 347 n.16
warp drives and negative mass objects 278–81
weakly interacting massive particles (WIMPs) 179–81, 194, 195, 222
  annihilation 208–11
Weyl Curvature Hypothesis (WCH) 129–30
Weyl tensor 102 n.1
Wheeler, J.A. 16, 249
Wheeler-DeWitt (WDW) approach 117, 134, 135

Wheeler-DeWitt equation 73 n.41
Wheeler-DeWitt quantum gravity approach 117
white dwarfs 176, 177, 180–2, 184, 222
  chemical evolution in 188–90
  evolution final phase 190–4
  powered by proton decay 186–8
  WIMP annihilation in 208–9, 211
Wick rotation 154
Witten, E. 292 n.21
WKB approximation 117, 119, 134
wormholes, traversable 275, 283, 284

Xin Zhang 40–1, 43, 73 n.36
X-property 324–5

Yates, J. 341

Zeh, H.D. 134, 135
Zenonian paradoxes 31, 243
zero-point energy 65–6, 69, 139–40
  ignoring of 139